Numerical Weather Prediction

On the benefits of multi-model ensembles:
idealized studies using the Lorenz 1963 model.

Forecasting Research Technical Report No. 492

Christine Johnson

email:nwp_publications@metoffice.gov.uk

©Crown Copyright

On the benefits of multi-model ensembles: idealized studies using the Lorenz 1963 model

Christine Johnson

Abstract

The international collaboration of forecast centres has recently allowed access to two-week ensemble forecasts from a variety of models. A key question is whether the calibration and combination of these forecasts can lead to significant improvements in probability forecasts.

Here, idealized experiments, based on the Lorenz 1963 model with time-dependent errors, are used to evaluate the benefits of a multi-model ensemble. In the first part, we investigate the impact of calibration on a single-model ensemble. The bias correction is successful in correcting errors at short lead times, but is sensitive to the calibration data at long lead times because flow-dependent biases dominate the forecast errors. As expected, the calibration is successful in improving the ensemble reliability but not the resolution.

In the second part, we investigate the benefits of a simple model combination and also more sophisticated methods including multiple-regression, skill-based weights and Bayesian Model Averaging (BMA). The simple model combination acts to improve the ensemble resolution and the model-dependent weights act to give further improvements in the ensemble reliability, with all three weight-estimation methods tested here performing equally well.

Contents

1 Multi-model ensembles **2**
- 1.1 Multi-model ensemble general framework . . . 4
  - 1.1.1 Bias correction . . . 4
  - 1.1.2 Combination . . . 5
  - 1.1.3 Variance adjustment . . . 6
  - 1.1.4 Multi-model Probabilities . . . 6
  - 1.1.5 Summary . . . 8

2 Idealized ensemble forecast system **8**
- 2.1 Lorenz models . . . 8
- 2.2 Ensemble forecasts . . . 12
- 2.3 Moving window approach . . . 13
- 2.4 Ensemble verification . . . 16

3 Ensemble calibration **18**
- 3.1 Impact of bias correction at different lead times . . . 19
- 3.2 Impact of the position of the window . . . 22
- 3.3 Use of reforecast data . . . 23

4 Ensemble combination **25**
- 4.1 Simple combination . . . 26
- 4.2 Weighted combination: skill-based . . . 31
- 4.3 Weighted combination: multiple-regression based . . . 31

4.4 Weighted combination: BMA . . . 32
4.4.1 BMA applied to individual ensemble members . . . 32
4.4.2 BMA applied to single-model ensemble means . . . 34
4.5 Model-dependent weights results . . . 37

5 Conclusions and discussion **37**

A Verification tools **38**
A.1 Ensemble Averaging . . . 38
A.2 Rank Histogram . . . 39
A.3 Reliability Table . . . 40
A.4 Brier Skill Scores . . . 41

B The EM algorithm **43**
B.1 MLE estimate . . . 44
B.2 MAP estimate . . . 44

1 Multi-model ensembles

In weather forecasting, it is common to base a prediction on a single deterministic forecast. Both the initial conditions and the forecast model contain errors, which lead to errors in the forecast. As the forecast lead time reaches the so-called limit-of-predictability, the forecast errors dominate the large-scale disturbances, so that little useful information can be gained from the forecast (Lorenz, 1963).

To be able to gain useful information from forecasts 15 days ahead, one must turn to probabilistic forecasts, where the prediction is based on an ensemble of forecasts, each with different initial conditions or a different forecast model. Such an ensemble gives a range of the possible outcomes, an indication of the most likely outcome and its associated uncertainty. In some instances, the ensemble trajectories may cluster together, showing an inherantly predictable situation; at other instances, the ensemble trajectories may diverge, showing a less predictable situation. These differences in the limits-of-predictability are illustrated with the Lorenz model by Palmer (1993). A further advantage is that the ensemble mean also gives a better forecast than a single deterministic forecast as the unpredictable components are filtered out of the forecast (Leith, 1974).

Ensemble prediction systems have been developed at many operational centres. For example, both the ECMWF and NCEP have produced operational ensembles since 1992 (Toth and Kalnay (1993) and Molteni et al. (1996)). Typically, an ensemble of initial conditions is generated by adding small perturbations to the analysis. An ensemble of forecasts can then be produced by applying the same forecast model to each initial condition. Sometimes, small variations may be added to the model to account for some of the model error, but it is still the variations in the initial conditions that dominate the statistics.

Poor-mans ensembles (Ziehmann, 2000) have also been suggested as ways to account for errors in both the initial conditions and the model. In this method, the deterministic forecasts from different operational centres provide the ensemble members so that the ensemble samples the uncertainties in both the initial conditions and models. Mylne and Robertson (2002) showed that the poor-mans ensemble gives better probabilities at short lead times than the single-model ensemble because the poor-mans ensemble takes a random sample of the analysis errors, rather than the non-random sampling methods designed for medium-range prediction. The disadvantage of the poor-mans ensemble is the

small ensemble size, which has a particularly large impact on the forecasting of extreme weather events that have low probabilities.

The aim of the multi-model ensemble (Evans et al., 2000; Palmer et al., 2004) is to create a relatively large ensemble by combining ensembles from different models. Multi-model ensembles take advantage of the availability of data from different operational centres, and account for the error in the model by combining forecasts that use different models, and account for the error in the initial-conditions errors by using ensembles of perturbed initial conditions. It has been shown in the context of seasonal forecasting (Hagedorn et al., 2005) that the combination of ensembles from different models results in more skill than the single ensembles considered separately, this improvement being not just due to the increased ensemble size, but from the information coming from different models.

Hagedorn et al. (2005) concluded that the key to the success of the multi-model concept lies in combining independent and skillful models, each with its own strengths and weaknesses. It is difficult to define the best single model; one model might be the best model in the summer whilst another model is the best in the winter. A multi-model is able to combine the strengths from all the different models. Evans et al. (2000) also concluded that "in order to benefit from combining ensembles, it is necessary (but not sufficient) that the component systems sample different populations".

Stephenson et al. (2005) presented a unified framework for forecasting. In the first step, data assimilation is used to find the probability of the initial model state given the observations. In the second step, ensemble forecasts are used to find the probability of the final model state given the initial model state. In the third step, so-called forecast assimilation is used to find the probability of a given forecast variable given the final model state. This step includes both forecast combination and calibration and is necessary because of the differences between the model state and reality.

Although a multi-model ensemble combines the strengths from different models, some models might be better than others at different times, and hence further improvements might be made by giving the models different weights. Previous research has show that model-dependent weights can give improvements, but care needs to be taken in how they are calculated and used. For example, Raftery et al. (2005) concluded that the BMA model-dependent weights created a better deterministic forecast, and Stefanova and Krishnamurti (2002) also showed an improvement from model-dependent weights, when considering probability forecasts from seasonal multi-model forecasts. However, Doblas-Reyes et al. (2005) concluded that model-dependent weights gave no significant improvement in the DEMETER ensemble, but there was a lack of robustness in the multiple-regression coefficients.

One of the benefits of combining forecasts from different models is that model errors can cancel out so that a multi-model ensemble is more reliable than a single-model ensembles. However, the impact of model errors in forecasts can also be reduced through forecast calibration. In calibration, the forecasts are adjusted so that the average statistical properties are similar to those of a reference data set. For example, if forecast temperatures are consistently too high, these can be reduced through bias correction. Thus, it may be the case that a calibrated multi-model ensemble has little benefit over a calibrated single-model ensemble. The results in Doblas-Reyes et al. (2005) suggested that both calibration and combination improve the ensemble predictions, although the results are sensitive to the reference data sets.

It is not clear how much improvement combination can bring over the improvement over calibration. Therefore, a key research question is to evaluate the benefits of a multi-model ensemble. This evaluation should determine the benefits of each of the components to the multi-model ensemble, including calibration, the simple combination of models and model-dependent weights. It may be that a calibrated-single-model ensemble is equally as skillful as a sophisticated combination of ensembles. If this is the case, a multi-model ensemble would not be worthwhile.

The evaluation of an operational system is time-consuming due to the vast amount of data needed.

Therefore it is appropriate to investigate these questions using a simple idealized system. Here, we choose to use the Lorenz 1963 model with seasonally dependent biases. This model is described in section 3, along with the verification tools that are used to evaluate the forecasts.

In section 2, the general framework for the calibration and combination of a multi-model ensemble is described. In this approach we consider explicitly the combination of forecast probability density functions (pdfs), in a similar way to Raftery et al. (2005). This is in contrast to the approaches suggested by Stephenson et al. (2005) and Doblas-Reyes et al. (2005) where only the combination of single forecasts were considered. In section 4, we evaluate the impact of ensemble calibration on a single-model ensemble. In section 5 we then evaluate the further improvements that can be obtained from a simple combination of ensembles, and then from the use of model-dependent weights. These results are followed by conclusions. Details about the verification tools and the computations for the Bayesian model (BMA) averaging are given in the appendices.

1.1 Multi-model ensemble general framework

The aim of ensemble forecasting is to produce forecasts of probabilities. Hence, it is in the framework of probabilities that the calibration and combination of ensembles is considered. It is assumed that the ensemble-mean is the best estimate of the forecast state. Indeed, one of the benefits of ensemble forecasting is that the ensemble mean is, on average, more accurate than a single deterministic forecast. We then choose to represent the single-model ensemble pdf by a normal-distribution with the mean given by the ensemble mean and the variance given by the ensemble-variance. Although the ensemble might have a different distribution, this normally distributed approximation is assumed to be good enough for the purposes of a post-processing framework and allows simple procedures to be applied.

The formation of the multi-model ensemble consists of three separate steps: bias correction of the ensemble-mean, calculation of model-dependent weights and ensemble variance adjustment. These steps are now described using the following notation.

We assume that there are M single-model ensembles, and that the k^{th} single-model ensemble has N_k ensemble members. The deterministic forecast of the k^{th} model is denoted by $\mathbf{x}_k^d$, and the i^{th} ensemble member for the k^{th} model is denoted by $\mathbf{x}_k^i$. The true model is denoted by $\mathcal{M}_t$ and the k^{th} model is denoted by $\mathcal{M}_k$.

1.1.1 Bias correction

The first step is to obtain the calibrated single-model ensemble-means. Although the ensemble-mean is more accurate than the deterministic forecast, it may still be affected by model biases. For example, the temperatures may be consistently too hot or have too large variations in time so that the warm periods are too warm and the cold periods are too cold. These biases may be reduced through the use of linear-regression. If the single-model ensemble mean from model $\mathcal{M}_k$ is denoted by $\bar{x}_k$, then the calibrated ensemble mean is given by

$$\bar{z}_k = \beta_0 + \beta_1 \bar{x}_k \tag{1}$$

where the regression coefficients β_0 and β_1 adjust the first and second moments respectively. Note that β_1 corrects the variance, over time, of the ensemble-mean, and not the spread of the ensemble. By minimizing the mean square error between $\bar{z}_k$ and forecasts from a calibration data set y, the equations for the coefficients are given by

$$\beta_1 = cov(\bar{x}_k, y)/cov(\bar{x}_k, \bar{x}_k) \tag{2a}$$

$$\beta_0 = \mu_y - \beta_1 \mu_x \tag{2b}$$

where $cov(\bar{x}_k, y)$ is the covariance of $\bar{x}_k$ with y and μ_x and μ_y are the means of $\bar{x}_k$ and y respectively. If the variance of $\bar{x}_k$ is too small, β_1 will be greater than one. If there is no correlation between $\bar{x}_k$ and y then β_1 is near-zero, because the predictor, $\bar{x}_k$, contains no useful information.

The bias coefficients, calculated using the ensemble mean, are applied to each member of the ensemble, along with the deterministic forecast. This results in an ensemble such that the corresponding ensemble-mean is well-calibrated. As noted by Hamill et al. (2004), using the ensemble-mean to calculate the regression coefficients exploits the ability of the ensemble averaging to filter out the smaller, unpredictable scales and retain the larger, predictable ones.

1.1.2 Combination

The second step is to combine the ensembles together using model-dependent weights. Rather than averaging the ensemble members from the single-models, we average the pdfs from the single-models in a similar way to Raftery et al. (2005).

From the law of total probability, this average can be written as:

$$p(x) = \sum_{k=1}^{M} p(x|\mathcal{M}_k)p(\mathcal{M}_k) \tag{3}$$

where $p(\mathcal{M}_k)$ is the probability of $\mathcal{M}_k$ being the best model. These probabilities can be viewed as model-dependent weights $w_k = p(\mathcal{M}_k)$ where $0 \leq w \leq 1$ and the weights sum to one, $\sum_{k=1}^{M} w_k = 1$. Methods to calculate these weights are discussed further on in this paper.

Although this equation forms the basis for the multi-model ensemble, it still does not describe how the ensemble can be formed. To simplify things, we approximate the bias-corrected single-model ensemble pdfs by normal distributions, f_k, with means $\bar{z}_k$ and variances σ_k^2, so that the combined pdf is a weighted average of the normal distributions for each model,

$$p(x) = \sum_{k=1}^{M} w_k f_k(\bar{z}_k, \sigma_k^2) \tag{4}$$

The average of the normal distributions gives the pdf for the multi-model ensemble, with mean and variance given by:

$$\bar{z}_{MM} = \sum_{k=1}^{M} w_k \bar{z}_k \tag{5}$$

$$\sigma_{MM}^2 = \sum_{k=1}^{M} w_k \sigma_k^2 + \sum_{k=1}^{M} w_k (\bar{z}_{MM} - \bar{z}_k)^2 \tag{6}$$

This last formula is known as the decomposition of variance (Draper, 1995; Raftery et al., 2005), as each model gives a different sample. The first term is the within-model variance, which describes the uncertainty of the estimate given different initial conditions but the same model; and the second term is the between-model variance, which describes the uncertainty of the estimate given different models, but the same initial conditions.

In summary, the multi-model combination is formed by the union of the ensemble members from each single-model, with each ensemble member being associated with a model-dependent weight, w_k. In a simple combination, the weight given to each model is identical, but in a more complicated combination, different weights are given to different models, based on past model performance. Methods to compute these weights are discussed later on in this paper.

1.1.3 Variance adjustment

The multi-model variance contains components due to both errors in the initial conditions (mainly described by the within-model variance) and in the model (mainly described by the between-model variance) and should therefore give a better estimate of the uncertainty compared with single-model ensembles or poor-mans ensembles. However, it may still be uncalibrated and so variance adjustment needs to be applied. It is important that this adjustment is applied after combination. If the single-model variances were adjusted before combination, then the variance of the multi-model would be too large, since model combination acts to increase the ensemble variance (with the between-model component) and to decrease the RMS errors of the ensemble mean.

Variance adjustment is achieved by adding a variance adjustment factor to each single-model ensemble (Doblas-Reyes et al., 2005). This is identical to the variance inflation that is applied in an ensemble Kalman filter (e.g. Whitaker and Hamill, 2002). Methods to calculate the values of the variance adjustment factor are discussed further on in this paper.

There are two components to the variance adjustment. First, the between-model variance can be adjusted. It is likely that the models may be too far apart so that the multi-model ensemble pdf is artificially bi-modal. In this case the between-model variance should be reduced by adding moving the single-model means closer to each other. The size of this reduction is controlled through the parameter γ.

$$\bar{z}_k^{\text{new}} = \bar{z}_k + \gamma(\bar{z}_{MM} - \bar{z}_k) \tag{7}$$

Second, the single-model ensemble variances, σ_k^2, should be adjusted so that the multi-model variance, σ_{MM}^2, is consistent with the RMS errors of the multi-model ensemble mean $\bar{z}_{MM}$. Typically a single-model variance is too small at long lead times because the model-error is not accounted for. In this case the variance adjustment factor, α_k, acts to inflate the ensemble.

The variance-adjusted, bias-corrected ensemble members are given by:

$$v_k^i = \bar{z}_k^{\text{new}} + \alpha_k(z_k^i - \bar{z}_k) \tag{8}$$

giving the variance-adjusted multi-model ensemble variance

$$\sigma_{MM}^2 = \sum_{k=1}^{M} w_k \alpha_k^2 \sigma_k^2 + \sum_{k=1}^{M} w_k \left((1-\gamma)(\bar{z}_{MM} - \bar{z}_k)\right)^2 \tag{9}$$

1.1.4 Multi-model Probabilities

Ensemble forecasts are often used to derive probabilities of given events, such as the probability of temperatures above $30 \circ C$. These remain straightforward to compute, since the multi-model ensemble probabilities are given by a weighted average of the probabilities associated with the calibrated single-model ensembles (Stefanova and Krishnamurti, 2002),

$$p_{MM} = \sum_k \left(\frac{1}{N_k} \sum_i w_k o_k^i \right) \tag{10}$$

$$= \sum_k w_k p_k \tag{11}$$

where o_k^i is the binary forecast that the event is observed, based on the i^{th} bias-corrected, variance-adjusted, ensemble member v_k^i from model k, and $p_k = \frac{1}{N_k} \sum_i o_k^i$ is the probability that the event is observed, based on model $\mathcal{M}_k$.

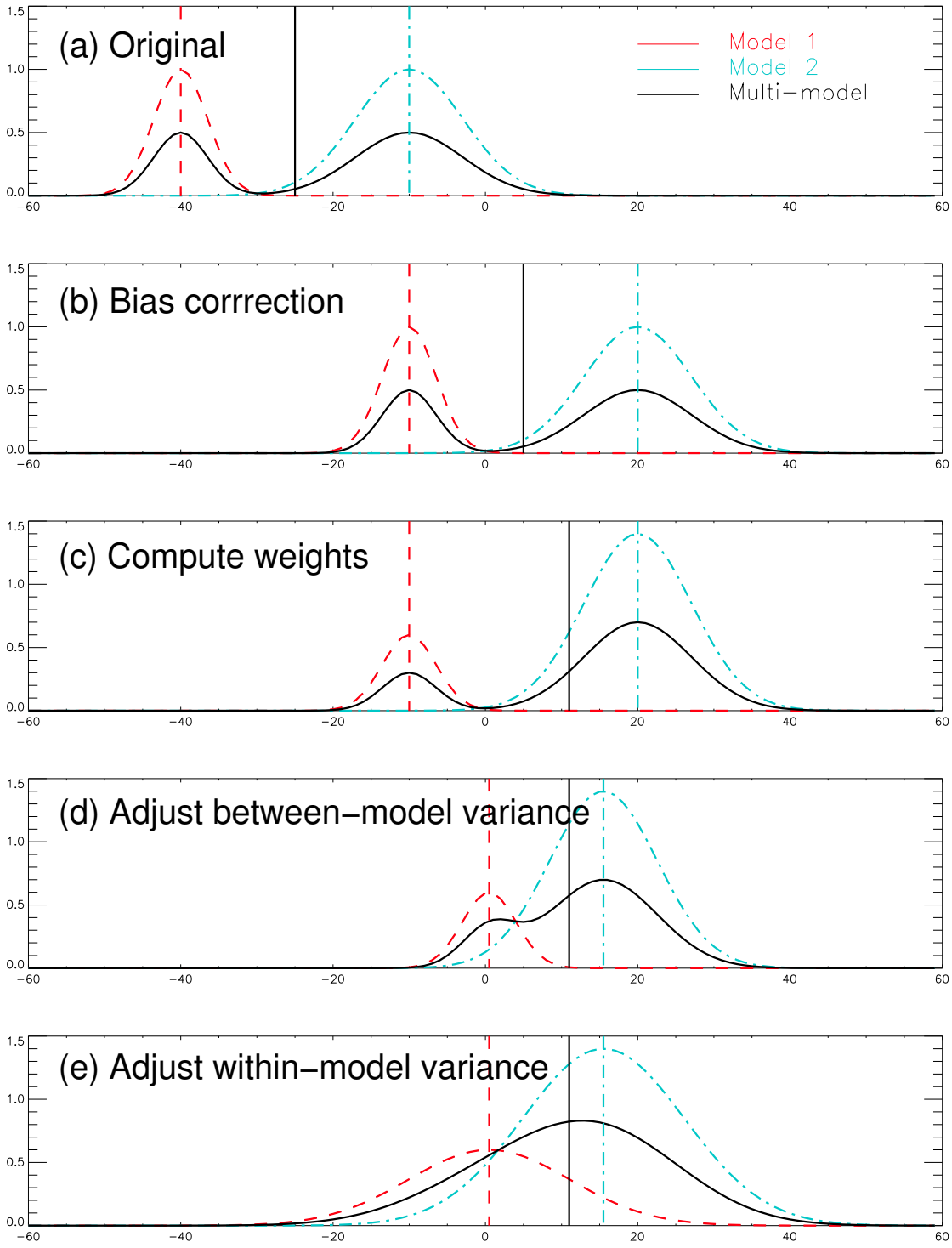


Figure 1: *Illustration of the steps in calibrating and combining the multi-model ensemble. Each panel shows the pdfs for each ensemble: the two single ensembles in red and blue and the multi-model ensemble in black. The ensemble means are shown by the vertical lines at the centre of each pdf.*

1.1.5 Summary

The multi-model ensemble is given by the union of bias-corrected single-model ensemble-members. Whilst the multi-model ensemble is a single entity, the identity of the model that each ensemble member is drawn from is still distinguished; the single-models are associated with model-dependent weights.

The steps to create the multi-model ensemble are illustrated by the pdfs in Fig.1. The original single-model ensembles are shown in panel (a). If we imagine that both ensembles have a negative bias, then the impact of the bias correction is to shift the pdfs along the x-axis, as shown in panel (b). The next step is to compute the model-dependent weights to give an indication of which model seems more likely. This adjusts the height of each pdf, as shown in panel (c). The final step is to adjust the variance of the multi-model ensemble. At first, because the two models sample different parts of phase space, the multi-model pdf is artificially bi-modal. By nudging each single-model towards the multi-model mean, the pdf becomes more uni-modal, as shown in (d). Finally, as shown in (e), the variances of the two single-models are adjusted to give the appropriate variance for the multi-model pdf.

2 Idealized ensemble forecast system

We now describe the design of the idealized ensemble systems that will be used to evaluate different techniques to combine and calibrate multi-model ensembles. The idealized framework is designed to mimic operational systems as closely as possible, but remains simple and easy to work with. Two forecast models are defined by adding seasonal bias terms so that the models have different levels of skill in different seasons. Analyses and ensemble members are created by adding random perturbations to the true state. The size of the model parameter errors and the initial condition errors are specified so that the growth rates of the RMS errors over a 15-day period are similar to those for operational systems.

2.1 Lorenz models

The three-variable Lorenz 1963 model (Lorenz, 1963) is used to define the 'true' state in all the following experiments. The model equations are:

$$\dot{x} = -\sigma(x - y) \tag{12a}$$

$$\dot{y} = \rho x - y - xz \tag{12b}$$

$$\dot{z} = xy - \beta z \tag{12c}$$

where $\sigma = 10$, $\rho = 28$, and $\beta = 8/3$. The model is integrated numerically using 4th-order Runge-Kutta time-stepping with a non-dimensional timestep of 0.005.

Examples of the forecasts produced by this model are shown in the top left panel of Fig. 2. In this z-x plane, the attractor resembles a butterfly with two wings. The state is sometimes found on the left wing, where $x < 0$, and sometimes on the right wing, where $x > 0$. These two wings are reminiscent of the regimes of behaviour found in the atmosphere. For example, in one month the atmosphere may be in a regime associated with rain over the UK, such as a sequence of depressions. Then a regime transition may occur so that in the next month the atmosphere is in a regime associated with dry weather, such a high pressure block.

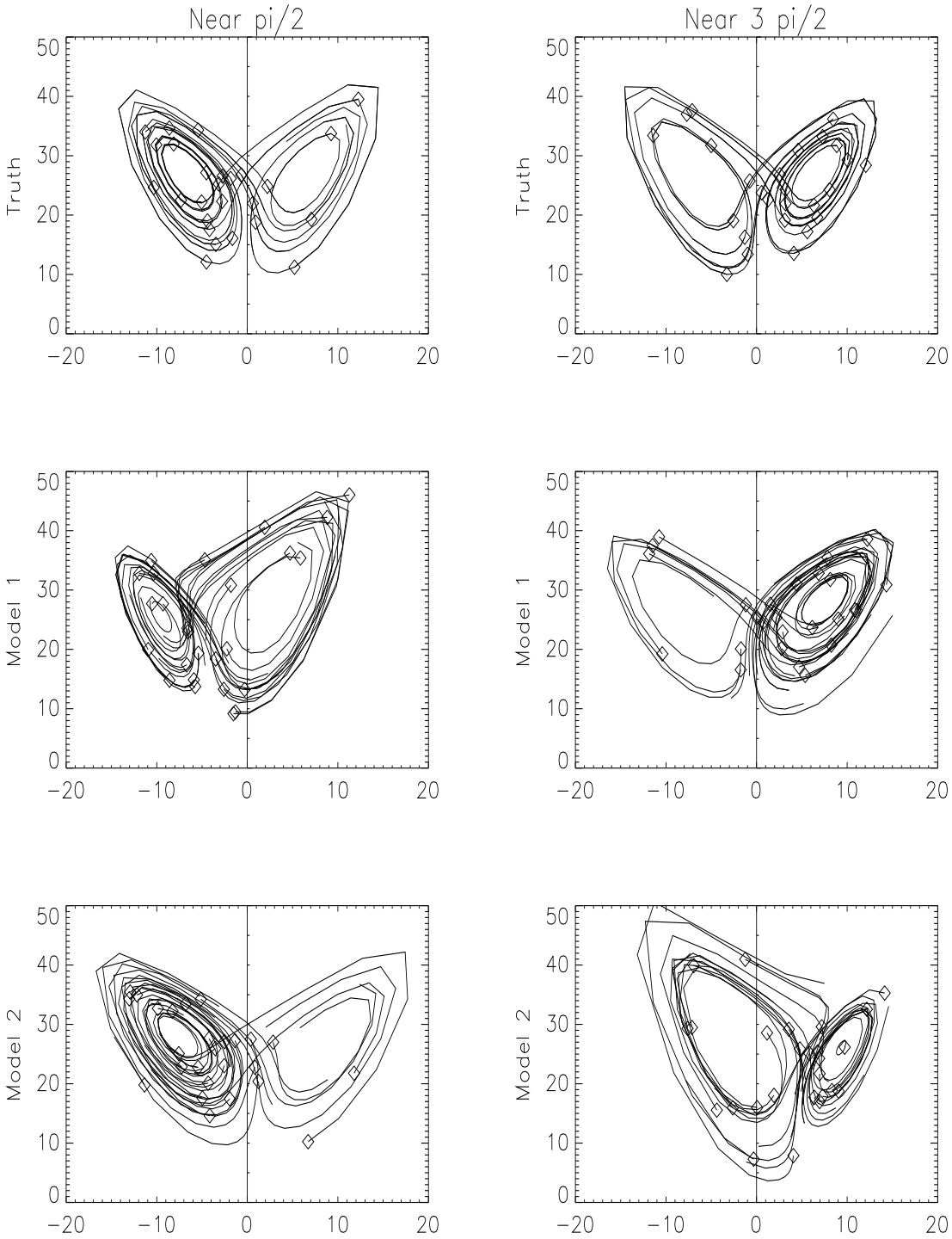


Figure 2: *The model attractors in two seasons defined by 6 day windows centred on $\theta = \pi/2$ and $\theta = 3\pi/2$. The attractors (z against x) for the truth (upper panel), model 1 (middle panel) and model 2 (lower panel). Each attractor is formed by a sample of forecasts from T+7 to T+15. The state at T+7 for each forecast is shown by the diamond.*

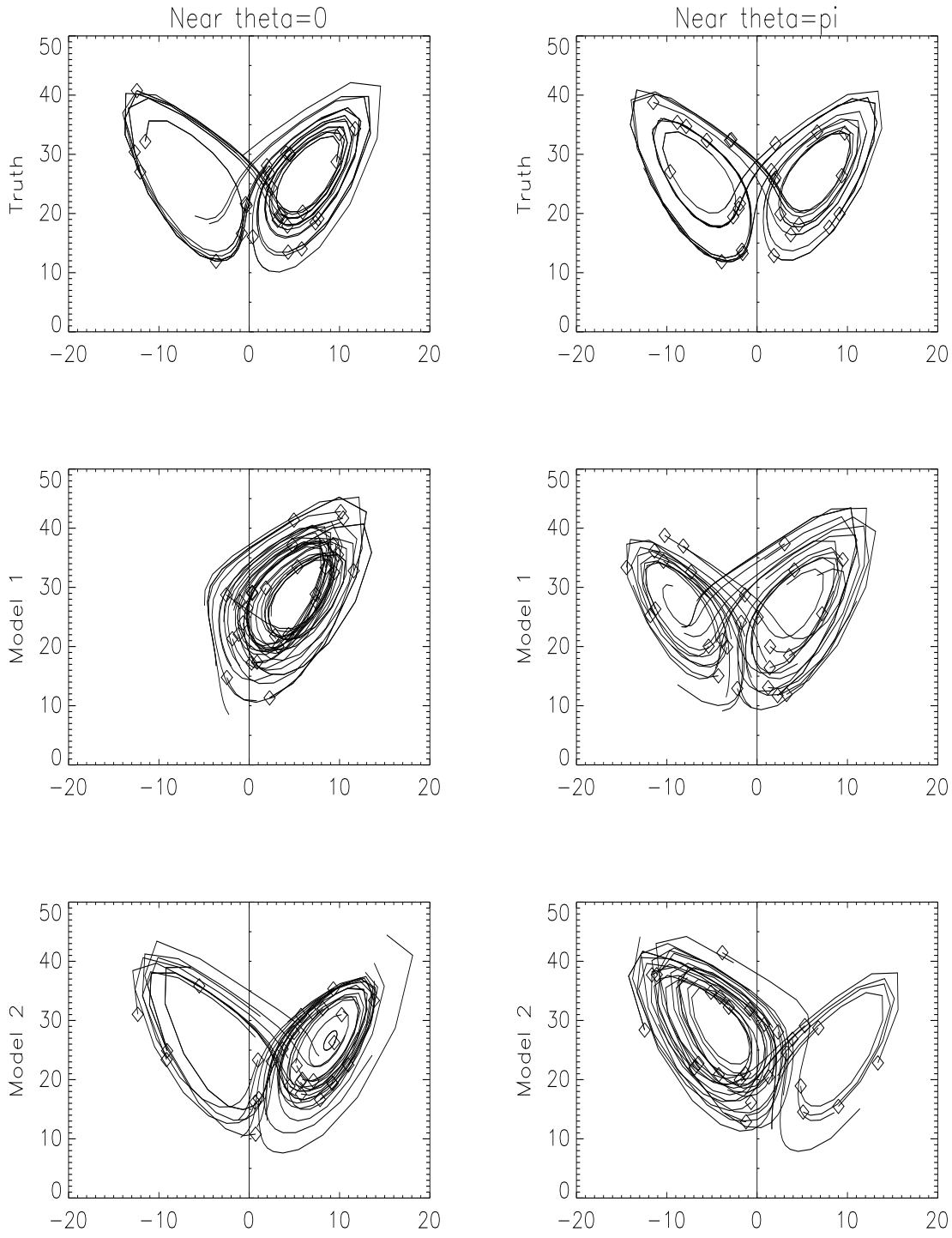


Figure 3: *The model attractors in two seasons defined by 6 day windows centred on $\theta = 0$ and Near $\theta = \pi$. (z against x) for the truth (upper panel), model 1 (middle panel) and model 2 (lower panel). Each attractor is formed by a sample of forecasts from T+7 to T+15. The state at T+7 for each forecast is shown by the diamond.*

The Lorenz model also exhibits the chaotic behaviour that results in the rapid growth of small errors, as seen in the atmosphere. Two forecasts with slightly different initial conditions can result in very different forecasts. Further, the local predictability properties depend on the position of the state on the attractor, so that sometimes a forecast is very accurate and at other times a forecast is inaccurate.

The existence of regimes and chaotic behaviour means that there are similarities between the Lorenz model and the atmosphere (Palmer, 1993). There are two timescales in the Lorenz model: the fast timescale describes the evolution of the state around the fixed point at the centre of each butterfly wing; the slow timescale describes the residence time in each wing as the state evolves and switches between the different regimes. Following Palmer (1993), we associate the fast timescale in the Lorenz model with the synoptic time scale in the atmosphere so that one nondimensional time unit in the Lorenz model corresponds to about 10 days in the atmosphere. This corresponds to a synoptic timescale of 5 days and a regime-residence timescale of a few weeks.

All forecast models in numerical weather prediction have errors that arise from numerical discretization, unknown model parameters and physical processes that are poorly represented or unknown. Here, we use the standard Lorenz model to define a true state, and two similar models are used as the forecast models. 15-day forecasts are produced at 12-hourly intervals over a two year period, where one model-year consists of 300 days.

A parameter θ is defined as

$$\theta = \frac{2\pi}{300} day. \tag{13}$$

θ varies between 0 and 2π over the year (300 days) and is used to define seasonal bias terms for the forecast models.

The idealized forecast models are identical to the standard Lorenz model, but with added seasonal bias terms, giving the equations

$$\dot{x} = -\sigma(x - y) - \eta_1\sigma + \eta_2 \tag{14a}$$

$$\dot{y} = \rho x - y - xz + \eta_1(\rho - z) - \eta_2 \tag{14b}$$

$$\dot{z} = xy - \beta z + \eta_1 y \tag{14c}$$

The parameter η_1 gives a bias term that acts to shift the whole attractor in the x direction. For model 1, $\eta_1 = 2(1 + \sin\theta)$ and for model 2, $\eta_1 = -2(1 - \sin\theta)$. Thus, when θ is close to $\pi/2$, η_1 produces a negative bias in model 1 and a negligible bias in model 2. Similarly, when θ is close to $3\pi/2$, η_1 produces a negligible bias in model 1 and a positive bias in model 2. This is illustrated by the attractors shown in the bottom panels of Fig. 2.

The parameter η_2 gives a bias term that acts to make the state spend more time on a particular wing of the attractor. For model 1, $\eta_2 = -8(1 + \cos\theta)$ and for model 2, $\eta_2 = 8(1 - \cos\theta)$. Thus, when θ is close to 0, model 1 spends more time on the right hand side of the attractor, whereas there is little change for model 2. Similarly, when θ is close to π, there is little change for model 1, but model 2 spends more time on the left side of the attractor. This is illustrated by the attractors shown in the bottom panels of Fig. 3.

This design means that these idealized forecast models exhibit different levels of skill in different seasons, in a similar way to real forecast models. In this situation, it is difficult to define the best single model, and hence a combination of the models should give an improvement in the forecast skill.

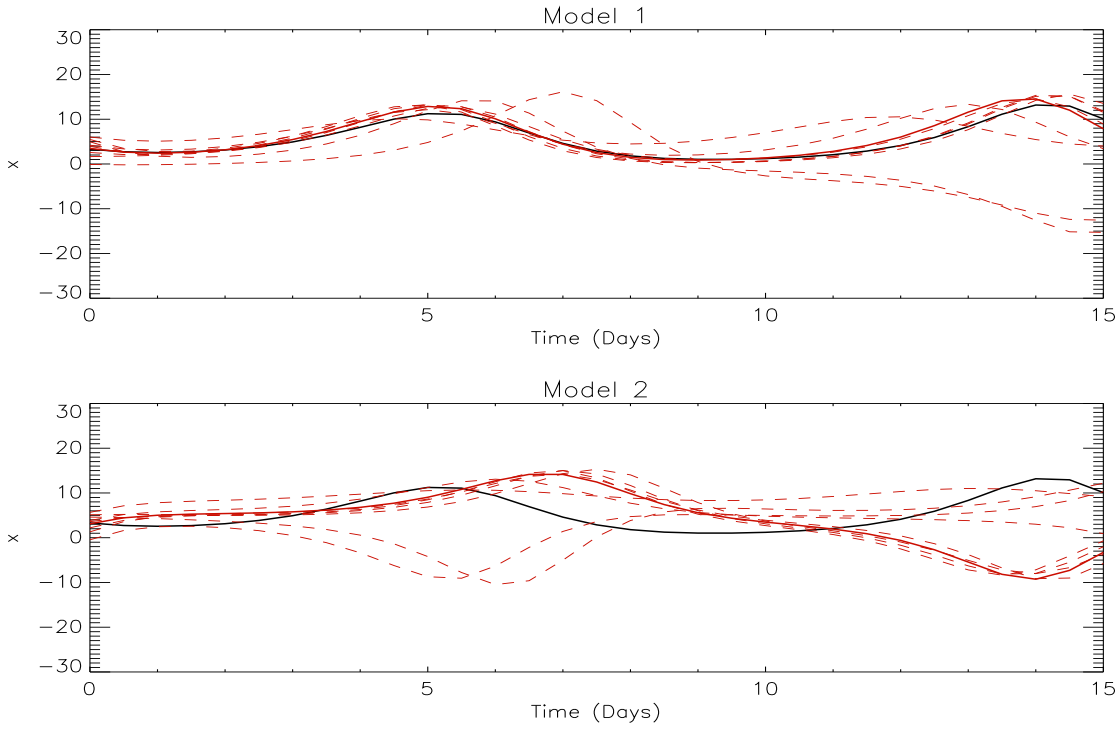


Figure 4: *The ensemble plumes on day 200 for model 1 (upper) and model 2 (lower), consisting of the truth (black solid), deterministic (red, solid), and ensemble members (red, dashed).*

2.2 Ensemble forecasts

In operational systems, the initial conditions for the deterministic forecast are given by the analysis which is produced from a complicated data assimilation system that blends previous forecasts with observations. Here we do not attempt to mimic the data assimilation system; it is sufficient to use initial conditions that have random errors. Operational systems generate an ensemble of initial conditions to sample the analysis errors in the best way. Methods such as singular vectors (ECMWF) or breeding vectors (NCEP) are popular as they draw from the directions that exhibit the largest error growth. Again, we do not attempt to mimic such systems and instead generate initial conditions by adding random perturbations to the analyses.

In the following experiments, we have $M = 2$ models, and each model has $N_k = 8$ members. The analysis (deterministic forecast at the initial time t_0) $\mathbf{x}_k^d(t_0)$ is given by the true state $\mathbf{x}^t(t_0)$ with an additive random analysis error $\boldsymbol{\varepsilon}_k(t_0) = (\varepsilon_x, \varepsilon_y, \varepsilon_z)$. The initial states for the ensemble members $\mathbf{x}_k^i(t_0)$ are then given by the analysis with perturbations, $\boldsymbol{\nu}_k^i = (\nu_x, \nu_y, \nu_z)$.

$$\text{Deterministic:} \qquad \mathbf{x}_k^d(t_0) = \mathbf{x}^t(t_0) + \boldsymbol{\varepsilon}_k(t_0) \tag{15a}$$

$$\text{Ensemble member:} \qquad \mathbf{x}_k^i(t_0) = \mathbf{x}^t(t_0) + \boldsymbol{\varepsilon}_k(t_0) + \boldsymbol{\nu}_k^i(t_0) \tag{15b}$$

The random analysis errors $\varepsilon_x, \varepsilon_y$ and ε_z are drawn from a Gaussian distribution with standard deviation 0.2. Similarly, the ensemble perturbations have a standard deviation of 2.0. These numbers are specified so that the spread of the ensemble is slightly larger than the RMS error for the first 5 days and slightly less than the RMS error for the last 5 days, as seen in operational ensembles.

The true state, deterministic forecasts and ensemble forecasts at the verification time $t_0 + t_f$ are

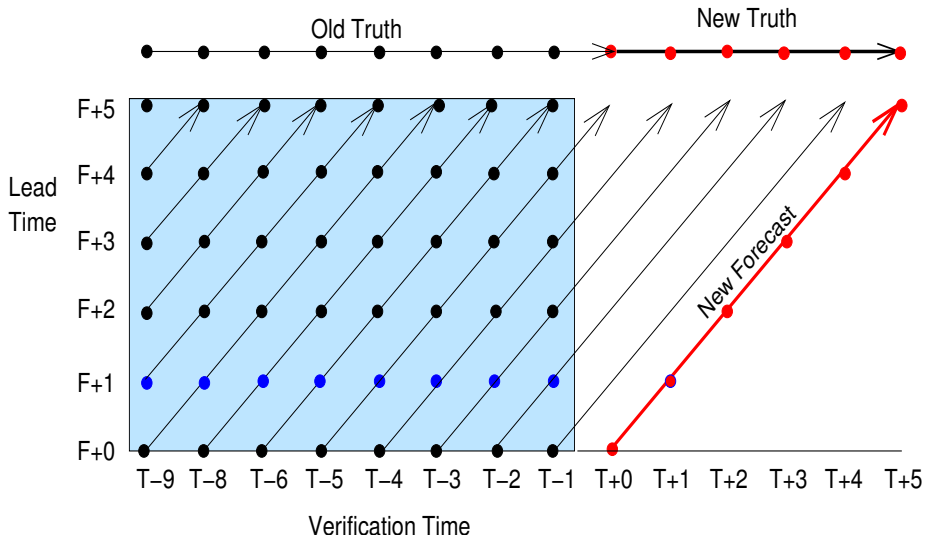


Figure 5: *Diagram of the moving-window postprocessing procedure. The arrows show individual forecasts starting at different times, with the red arrow showing the forecast to be calibrated. The blue box indicates the forecast values that can be used with the 'Old Truth' values to calculate the postprocessing parameters. For example, the eight forecasts at a lead time of F+1, from a range of verification times, are used to calibrate the new forecast valid at T+1.*

then produced by applying the appropriate model to the set of initial conditions.

$$\text{Truth:} \qquad \mathbf{x}^t(t_0 + t_f) = \mathcal{M}_t(\mathbf{x}^t(t_0)) \tag{16a}$$

$$\text{Deterministic:} \qquad \mathbf{x}_k^d(t_0 + t_f) = \mathcal{M}_k(\mathbf{x}_k^d(t_0)) \tag{16b}$$

$$\text{Ensemble member:} \qquad \mathbf{x}_k^i(t_0 + t_f) = \mathcal{M}_k(\mathbf{x}_k^d(t_0) + \boldsymbol{\nu}_k(t_0)) \tag{16c}$$

The models are integrated to produce 8-member, 15-day ensemble forecasts for each model, every 12 hours. The number of ensemble members and forecast length is therefore similar to operational ensemble systems. The forecasts are archived at 12-hourly intervals.

Examples of the ensembles from the two single models are seen in Fig. 4. The true state (black, solid) illustrates the synoptic timescale of about 5 days. The deterministic forecasts for models 1 and 2 (red, solid) start from slightly different initial conditions (to represent the analysis error). The deterministic forecast is closer to the true state for model 1 than model 2. In particular, near the end of the forecast, model 2 switches regime to the left side of the attractor instead of remaining of the right side. The ensemble members (red, dashed) start from different initial conditions again, that surround the deterministic forecast. The spread of the ensemble members around the deterministic forecast tends to increase with time.

2.3 Moving window approach

An important part of calibration and combination is the sample of data that is used to calculate the parameters. Typically, in other areas of research, there is a large amount of data to perform the calibration. However, in ensemble post-processing, there is a limited amount of data. This is due to the fact that models are constantly being improved and hence the statistical properties of the forecasts are changing. It is also due to the fact that in weather forecasting, we have no knowledge of the truth in the future. Hence, calibration and combination must always be based on forecasts from previous times.

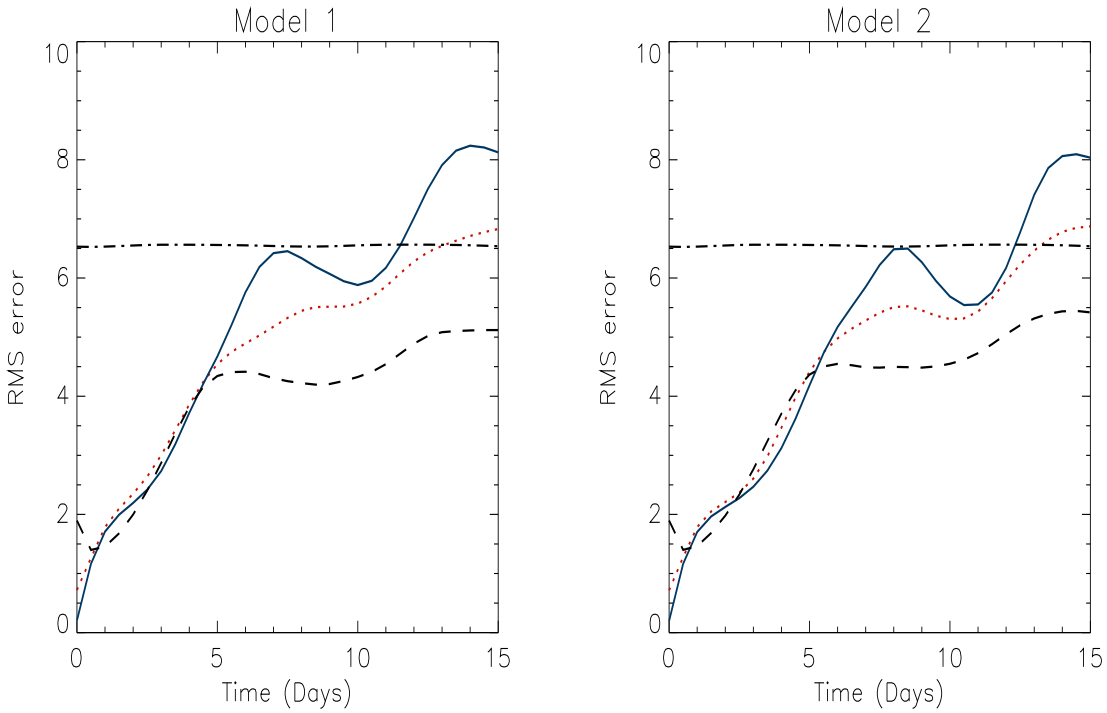


Figure 6: *RMS errors of the deterministic forecast (blue solid), ensemble mean (red dot) and climate (black dash-dot) as a function of lead time. Also shown is the ensemble spread (black dash), which is the standard deviation of the ensemble members around the mean.*

In a moving window, the sample of data used to calibrate the ensemble at a single moment in time is given by the previous p forecasts. An example of such a procedure is shown in Fig. 5. The forecasts in the blue shaded box, from T-9 to T-1, are used to calibrate the new forecast starting at T+0. The new forecast is then verified by the new truth. The window of forecasts is then cycled so that the forecasts from T-8 to T+0 are used to calibrate the new forecast starting at T+1. The forecast that starts, for example, at T-2 contains only two values that can be used to calibrate the new forecast at (T+1,F+1). The remainder of the forecast can not be used as there are no true values available to use in the calibration.

In the examples that follow, we use the previous 100 forecasts (50 days) in the moving window. This value is large enough to reduce sampling errors, and small enough so that the seasonality of the errors can be extracted. The true state is used as the calibration data. In a realistic case, the true state would not be available to calibrate against. Instead, model analyses or observations would need to be used.

In all the following methods, we perform the calibration and combination at each lead time separately so that correcting the forecast at 15 days is completely separate from correcting the forecast at 10 days. This method means that we cannot ensure that the calibrated forecast at 15 days is meteorologically consistent with the calibrated forecast at 10 days. However, it is likely that they will be reasonably consistent since the data used to calculate the calibration parameters will be consistent.

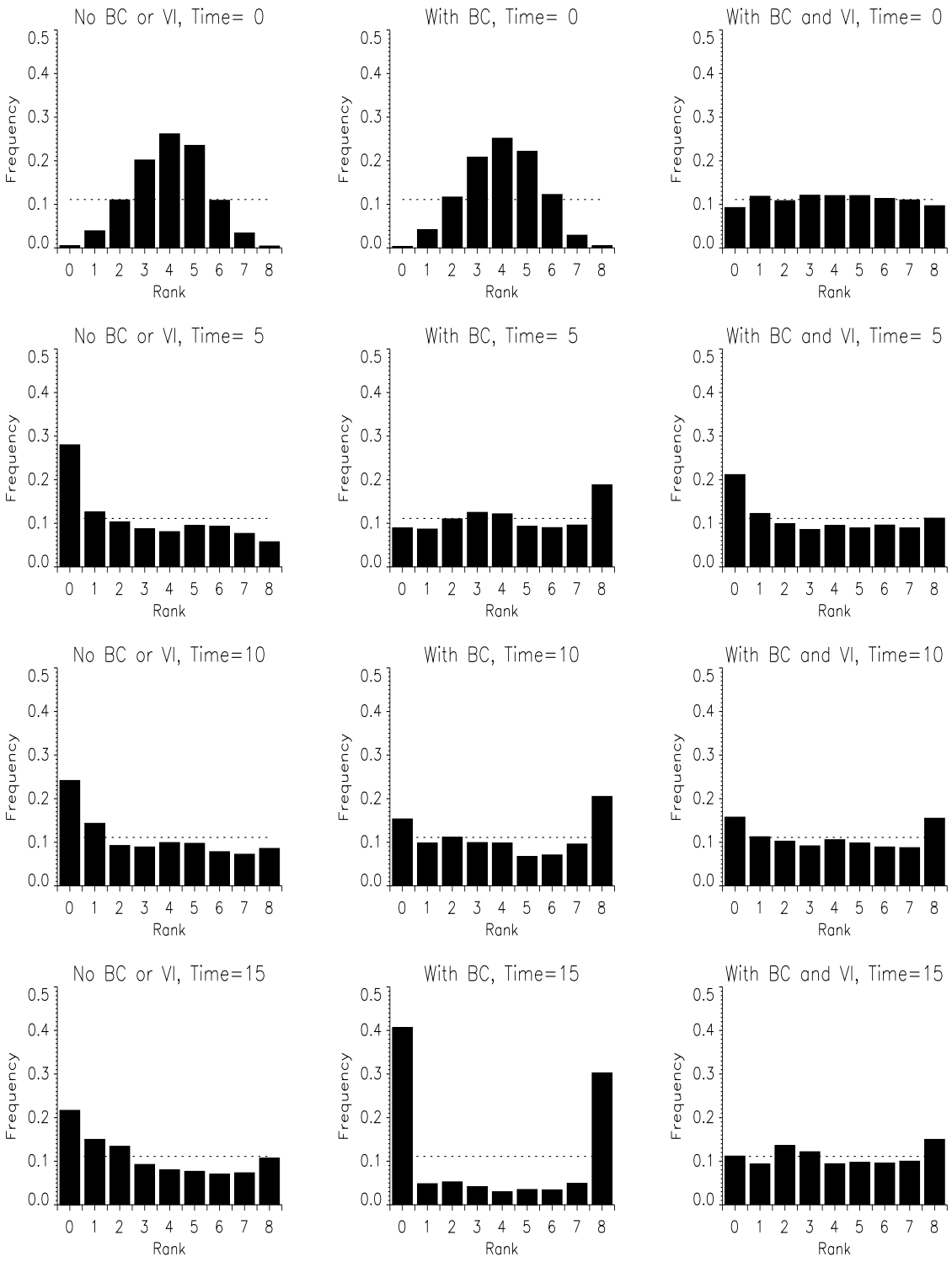


Figure 7: *Rank histograms for model 1 at lead times of 0,5,10 and 15. The left hand panels show the results with no bias correction or variance inflation, the centre panels show the results with bias correction only, and the right hand panels show the results with both bias correction and variance inflation.*

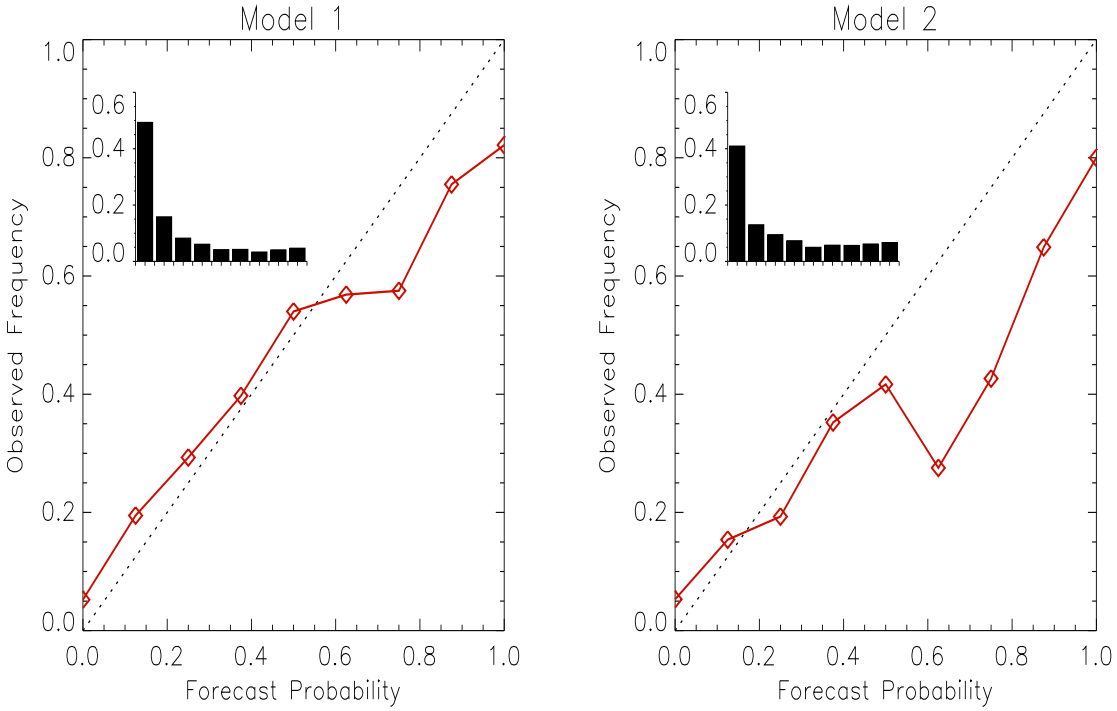


Figure 8: *Reliability diagrams ($p(o_l = 1|q_l)$ against q_l) for the ensembles of the single models. The perfectly reliable forecast is shown by the dashed line. The sharpness diagram (bar graph of $p(q)$ against q) is superimposed on each diagram. The event considered is whether $x > 6$ at a lead time of 7 days.*

2.4 Ensemble verification

We now use a basic set of verification measures to evaluate the skill of the single-model ensembles. A standard selection of verification measures are used to evaluate the ensembles and hence they are not described here in detail. However, details concerning the computation of the measures can be found in Appendix A.

In ensemble verification, it is necessary to select a set of variables to verify. These variables should be selected on the basis of whether they have a high user value Palmer (2002) or whether it gives a good representation of the attributes of the forecast (Stanski and Wilson, 1989). In numerical weather prediction, it is often difficult to find a single variable that fully assesses the overall skill of the system, although the 500mb height or mean sea level pressure is often used. In these idealized experiments with the Lorenz model, it is possible to verify only one variable. The two regimes of behaviour of the Lorenz model can only be seen in the x and y variables, and not z. Further, the x and y variables have a significant correlation, and so it is only necessary to verify the x variable. For the remainder of this paper, we only discuss the calibration, combination and verification of the x variable.

The RMS errors for the ensembles from the two models are shown in Fig.6. The RMS error of the ensemble mean is similar to the RMS error of the deterministic forecast for the first 5 days, but eventually, as the effects of nonlinearity begin to dominate, the errors in the deterministic forecast grow to become larger than those in the ensemble mean. The RMS errors for the deterministic forecast and the ensemble-mean are similar to that of the climate RMS errors by the end of the forecast. This is very similar to the results produced for operational systems (compare with Fig. 7.6 in Toth et al. (2003), where the ensemble mean error has reached the climatological RMS error by day 15). Initially the ensemble spread is similar in size to the ensemble mean error, but the growth of the ensemble spread

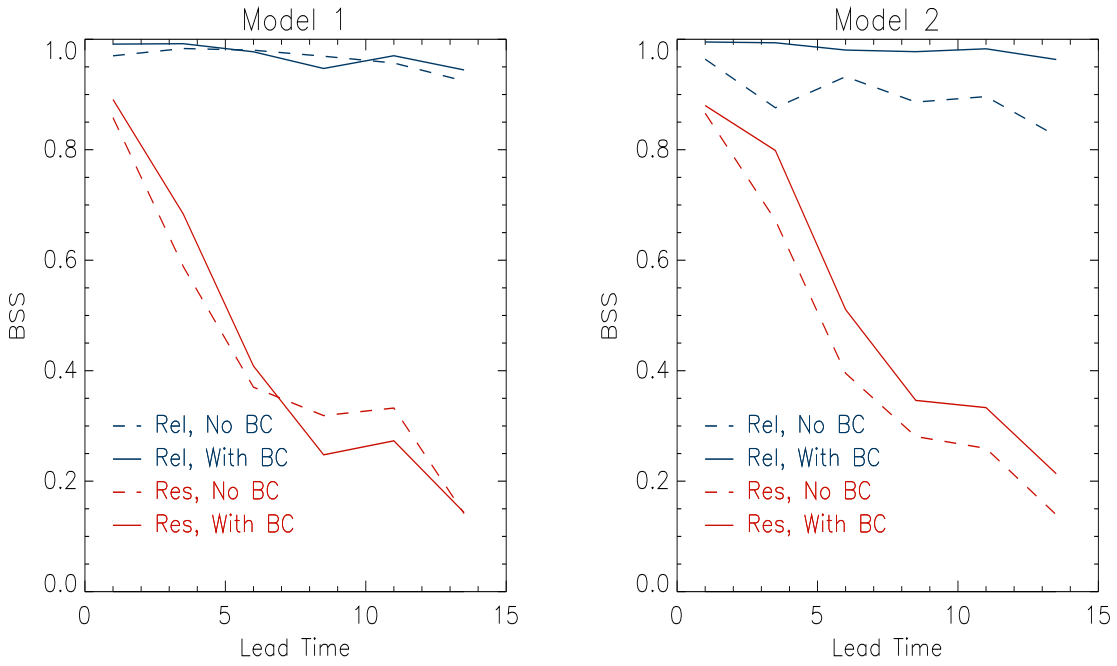


Figure 9: *Evolution of the Brier skill score components (reliability (blue) and resolution (red)) shown as a function of lead time. The event considered is whether $x > 6$. The solid curves show the values for the raw data, whereas the solid curves show the values for the data after bias correction and variance adjustment.*

is less than that of the error so that the ensemble spread is smaller than the ensemble mean error at the end of the forecast. Again, this behaviour is typical of operational forecast systems because they do not fully account for the effects of model error.

It is not only important to verify the ensemble spread, but to verify whether the true state lies within the ensemble. This can be assessed using the rank histogram, which is shown for model 1 in the panels on the left of Fig.7. At T+0, the histogram is overpopulated in the centre, showing that the ensemble spread is too large. At longer lead times, it is the smallest ranks that are overpopulated due to the bias in model 1.

Ensemble forecasts also need to be verified in a probabilistic sense. There are two aspects to the skill of a probabilistic forecasting system. The first is the **reliability**, which is the ability to provide unbiased estimates of the observed frequencies associated with different forecast values. The second is the **resolution**, which is the ability to discriminate between situations that lead to different forecast events in the future. An example of these two terms is to consider a thermometer. If the thermometer tells you when it will be hot and cold, but is uncalibrated so that it does not the right values of the temperature, it has a good resolution but a poor reliability. By simply relabelling the markings on the thermometer, we would get perfect accuracy. If the thermometer gives the average temperature all the time, then it has a good reliability but a poor resolution. Ideally, we would want the thermometer to have both a good resolution and reliability. It is the same idea for ensembles, except that the ensembles are dealing with probabilities instead of temperatures.

There are a variety of methods to assess the resolution and reliability aspects to ensemble forecasts. One method is the reliability diagram, which compares the observed frequencies of an event with forecast probabilities (see Appendix A for further details). Fig. 8 shows that both models have a poor resolution

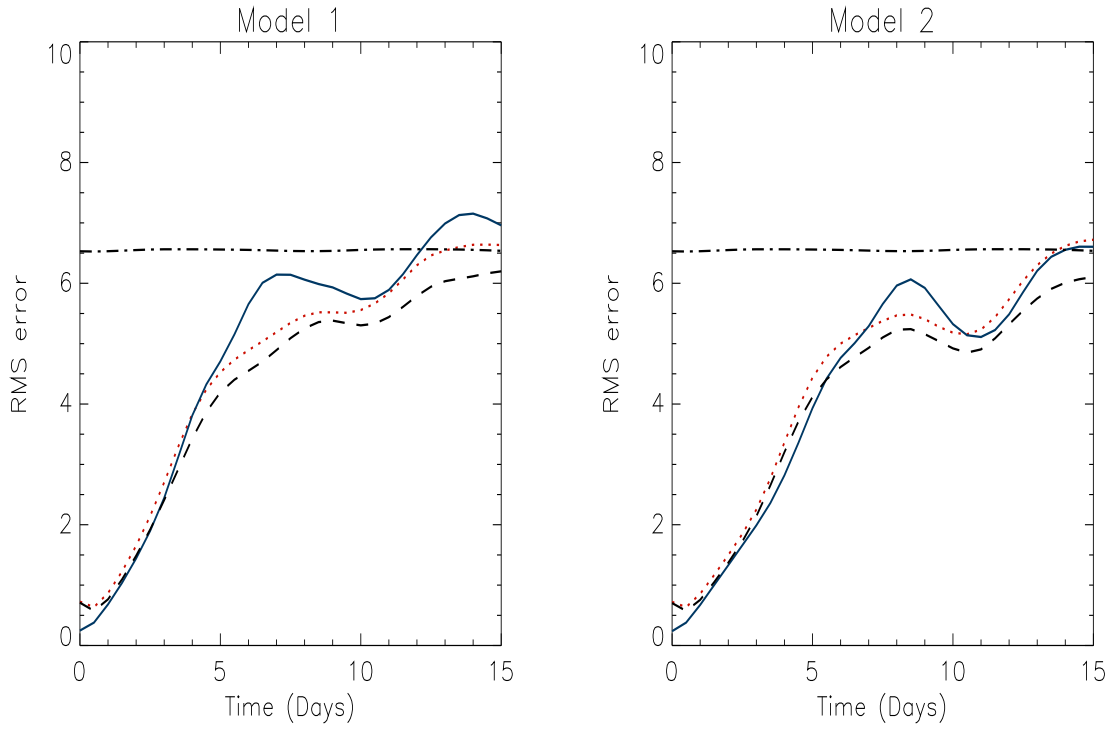


Figure 10: *RMS errors for models 1 and 2, with bias correction and variance inflation. The details are as for Fig. 6.*

and that the ensemble for model 2 has an overforecasting bias due to the positive bias in model 2. The reliability diagram shows only a single time instance and requires a subjective interpretation. However, it can be translated into quantitative values through the use of Brier Skill Scores, described in Appendix A and shown in Fig. 9. The dashed lines show that both the reliability and resolution decrease with increasing lead time.

3 Ensemble calibration

The purpose of ensemble calibration is to use a calibration data set to adjust the ensemble forecasts so that they are more skillful. The aim of this section is to evaluate the impact of calibration on a single-model ensemble. This includes both the impact of bias correction and of variance adjustment.

We first evaluate the impact of bias correction on model 1 using rank histograms. In Fig. 7, the left hand panels show that the ensemble spread is too large at T+0 and that at longer lead times there is a negative bias. The middle panels show the rank histograms after bias correction has been applied. The bias correction is applied to the ensemble mean (and hence also the ensemble members) using the linear regression formulae (equations 1 and 2a). The histograms are now longer overpopulated in the smallest ranks, but the extreme ranks are now overpopulated due to the ensemble underdispersion. This underdispersion is common in many operational ensembles as a result of the effects of model error being neglected.

We next evaluate the impact of variance adjustment, which is applied using (8). The parameter α_k is chosen so that the ensemble variance is representative of the ensemble-mean MSE,

$$\alpha_k^2 = MSE/var \tag{17}$$

where

$$MSE = \frac{1}{T}\sum_{j=1}^{T}(\bar{z}(t_j) - x^t(t_j))^2 \tag{18}$$

is the mean-square-error of the ensemble-mean over the calibration period $[t_1, t_T]$ and x^t is the true state.

$$var = \frac{1}{T}\sum_{j=1}^{T}\frac{1}{N}\sum_{i=1}^{N}(z^i(t_j) - \bar{z}(t_j))^2 \tag{19}$$

is the ensemble-variance over the calibration period. The rank histograms, shown in the right hand panels of Fig. 7, are now much more similar to the ideal horizontal line (dashed). The bias correction has reduced the overpopulation in the smallest ranks, and the variance adjustment has reduced the overpopulation of the extreme ranks at long lead times.

Comparing the RMS errors of the uncalibrated ensembles (Fig.6) with those of the calibrated ensembles (Fig. 10), we see that the bias correction acts to give reductions in the RMS errors of the deterministic forecasts at all lead times, reductions in the RMS errors of the means at days 0-3, and increases in the ensemble spreads at long lead times, so that there is consistency between the ensemble spread and the RMS errors. It is disappointing to see that although the bias correction has removed the bias of the ensemble, as shown in the rank histogram, it has not improved the RMS errors at long lead times.

Comparing the Brier skill scores of the uncalibrated ensemble with those of the calibrated ensembles (Fig. 9), we see that for model 1, the reliability is similar, and the resolution is improved at short lead times and worsened at long lead times. In contrast, for model 2 there is an improvement in both the reliability and resolution. The reason for the differences in the resolution between the two models is because the models have biases of different sign. The negative bias in model 1 means that when bias correction is applied, the average probability of the (rare) event $x > 6$ is larger. In this sense, the ensemble is less sharp, giving a worse resolution. The positive bias in model 2 means that when bias correction is applied, the average probability of $x > 6$ is smaller and hence the ensemble is more sharp, giving a better resolution. This argument is confirmed by Brier skill scores for the event $x > -6$ (not shown). In this case, for model 1, there is an improvement in both the reliability and the resolution, whereas for model 2 there is little change in the reliability and the resolution is worse. Thus the bias correction acts to improve the overall reliability but with no change to the overall resolution at different thresholds.

3.1 Impact of bias correction at different lead times

The bias correction only improved the RMS errors at short lead times and not at long lead times. To understand these differences, we examine the bias coefficients that are calculated by the linear regression. Fig. 11 shows the coefficients for both models at three different lead times as a function of the verification time. At T+1, β_1 is close to unity for both models at all verification times, and β_0 varies sinusoidally such that when model 1 has a negative bias (and hence positive bias coefficient), model 2 has a negligible bias and when model 2 has a positive bias, model 1 has a negligible bias. By T+5, the behaviour of both coefficients is more irregular in time. This is emphasised even more by T+15, by which point the forecasts are also showing less skill, as indicated by the near-zero values of β_1.

The reason that the bias coefficients fluctuate rapidly at longer lead times is the interaction of model errors with initial conditions. The bias coefficients for model 1 are shown as a function of both

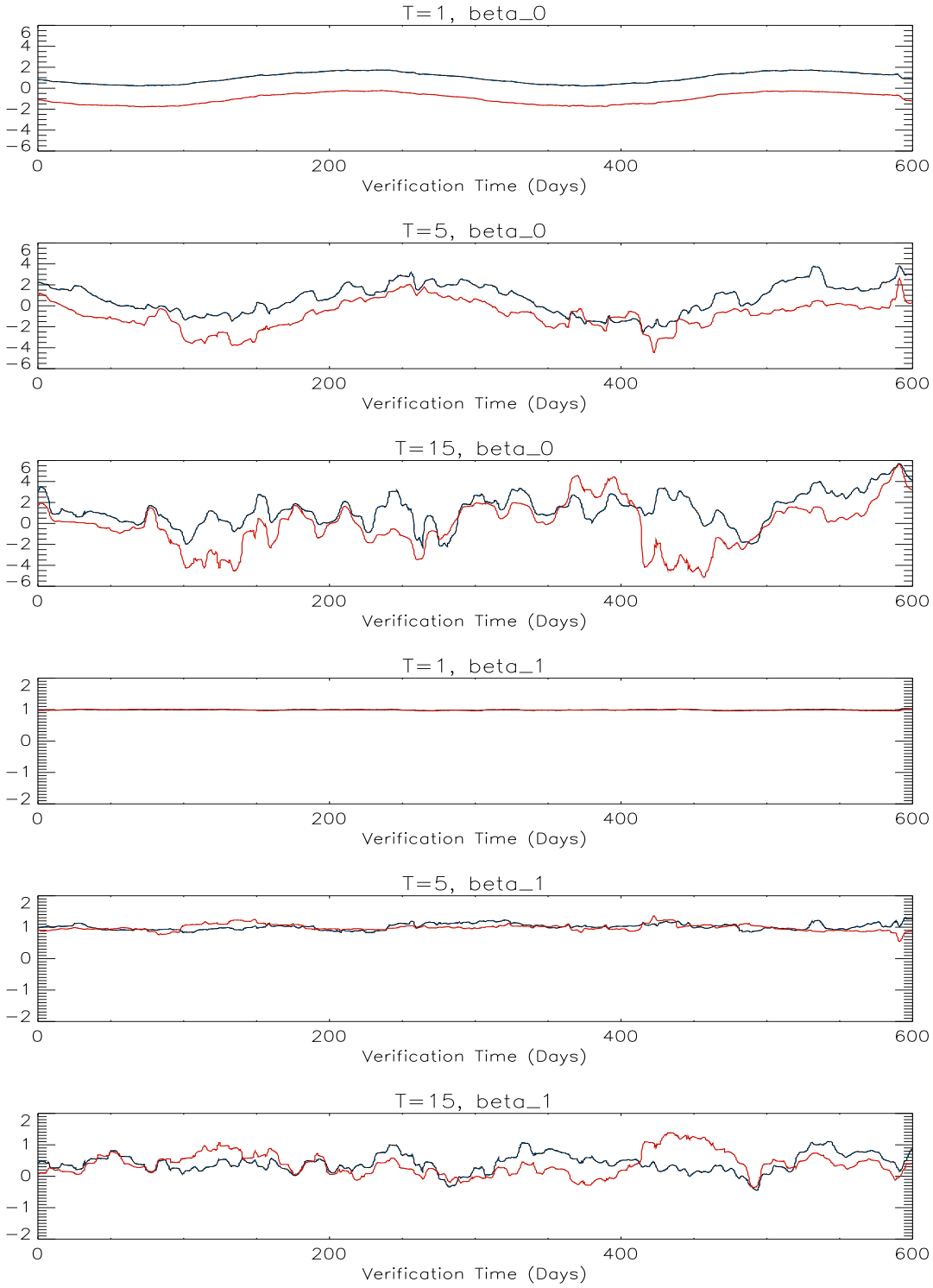


Figure 11: *The first and second bias coefficients over a 2 year period that starts with $\theta = \pi$. (600 days) for three lead times: T+1, T+5 and T+15 for model 1(blue, dash-dot) model 2(red, solid).*

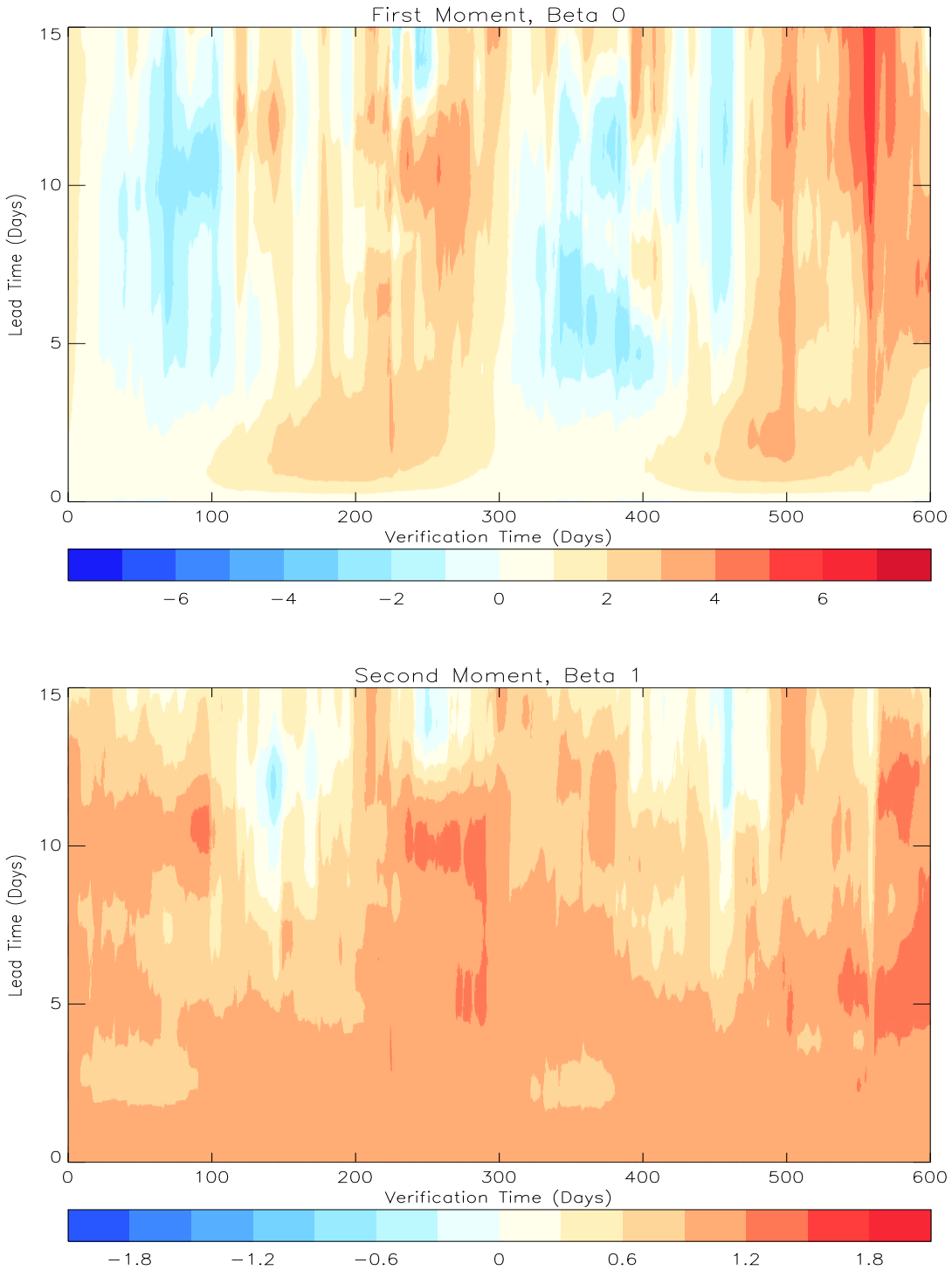


Figure 12: *The bias coefficients for model 1 as a function of lead time and verification time.*

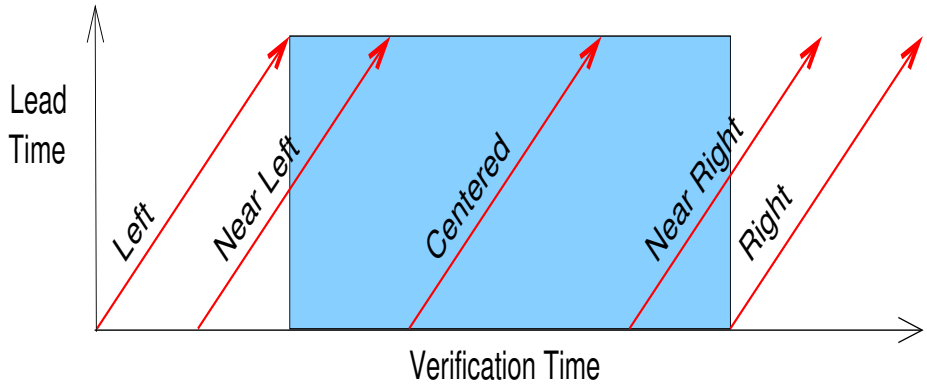


Figure 13: *Bias Correction frameworks. The box shows the window of forecasts used to generate the calibration parameters, and the diagonal, red arrows shows the relative position of the forecast to be calibrated.*

lead time and verification time in Fig.12. The sinusoidal seasonal bias can clearly be seen in β_0 at short lead times. As the forecast progresses, the model biases interact with the initial conditions to give 'flow-dependent' biases. These biases grow in time so that they eventually dominate over the slowly varying model bias. As the initial conditions vary rapidly over time, the forecast bias at longer lead times also varies rapidly over time.

The flow-dependence of the biases is confirmed by an identical experiment but with a different set of initial conditions for the true state. The bias coefficients (not shown) are similar to the previous bias coefficients with a smooth sinusoidal variation at T+1 and more rapid fluctuations at T+15. However, the positions of the maxima and minima in the fluctuations are now at different points in time. This is due to the differences in the initial conditions.

3.2 Impact of the position of the window

The fact that the biases at long lead times vary rapidly in time suggests that they are more sensitive to the calibration data that is used in their calculation. Therefore, we now investigate the impact of the position of the calibration window relative to the calibrated forecasts.

Fig. 13 shows five different frameworks with the calibration window in a different position relative to the verifying forecast. In the 'right' framework (Fig. 13a) the previous forecasts are used to calculate the calibration parameters. This is the only framework that is possible in reality, because in all other frameworks the truth is needed before it has actually occurred. In the 'centred' framework, (Fig. 13b) both forecasts before and after the calibrated forecast are used. In the 'left' framework, the window of forecasts is after the verifying forecast.

The RMS errors of the ensemble means using the 'right' and 'left' frameworks are shown in the left panel of Fig.14. For both models, both the 'left' and 'right' frameworks reduce the RMS errors in the 0-3 day period. The RMS errors at other lead times are virtually unchanged. This difference is because at short lead times, the bias coefficients have little variation in time, and hence are not as sensitive to the calibration data.

The RMS errors at days 6-14 are also reduced further if the verifying forecast is moved from the 'left' further into the window to the 'near left' position. Similarly, the RMS errors at days 3-8 are reduced if the forecast is moved further into the window, from the 'right' to the 'near right' position. The RMS errors at every lead time are reduced if the calibrated forecasts are in the centre of the calibration windows. This is because a window centred on the forecast gives a sample with forecast errors that have the most similar properties to the forecast to be calibrated. It is interesting to note

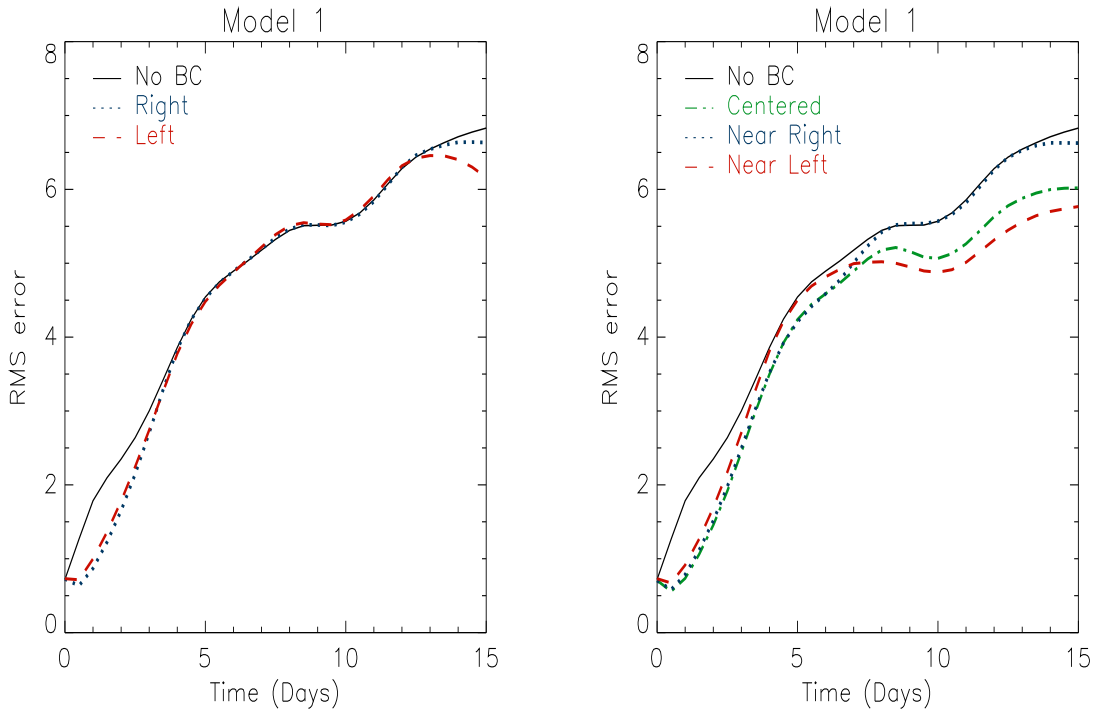


Figure 14: *The RMS errors of the single model ensemble means with and without bias correction for different calibration windows. No BC refers to the cases without bias correction. The other lines refer to the cases illustrated in Fig. 13.*

that the errors at the longest lead times are corrected the most using the 'near-left' window position, although this position gives little improvement at shorter lead times.

To summarize, at long lead times, the calibrating window needs to be close to the forecast so that the appropriate bias correction coefficients are calculated. At short lead times, the fluctuation of the bias is negligible and hence the position of the calibration window does not affect the calibration of short lead times as much.

3.3 Use of reforecast data

It is not possible in reality to use such a 'centred' moving window. However, it is possible to use a centred window based on reforecast data instead. Reforecast data is a series of forecasts from the same model, over many years, based on reanalysis data (Hamill et al., 2004). Using 23 years of data, Hamill et al. (2004) found that the week-2 calibrated forecasts were more skillful than the operational 6-10 day forecasts and that most of the benefit could be achieved with 10 years of training data.

We now investigate the use of reforecast data in the Lorenz model context. Following Hamill et al. (2004), the training dataset consists of 50 days of data over 20 years, centred in time about the date of the forecast to be calibrated. The 20 years of reforecasts are initialised from random initial conditions and the same model is used to produce the forecasts.

The bias coefficients for the two models are shown in Fig. 15. The cofficients at T+1 are similar to those for the moving window, with a sinusoidal variation. The coefficients at T+5 and T+15 also show a smooth sinusoidal variation; the flow-dependent fluctuations that are dependent on the initial conditions have been averaged out, leaving only the seasonally-dependent component of the bias.

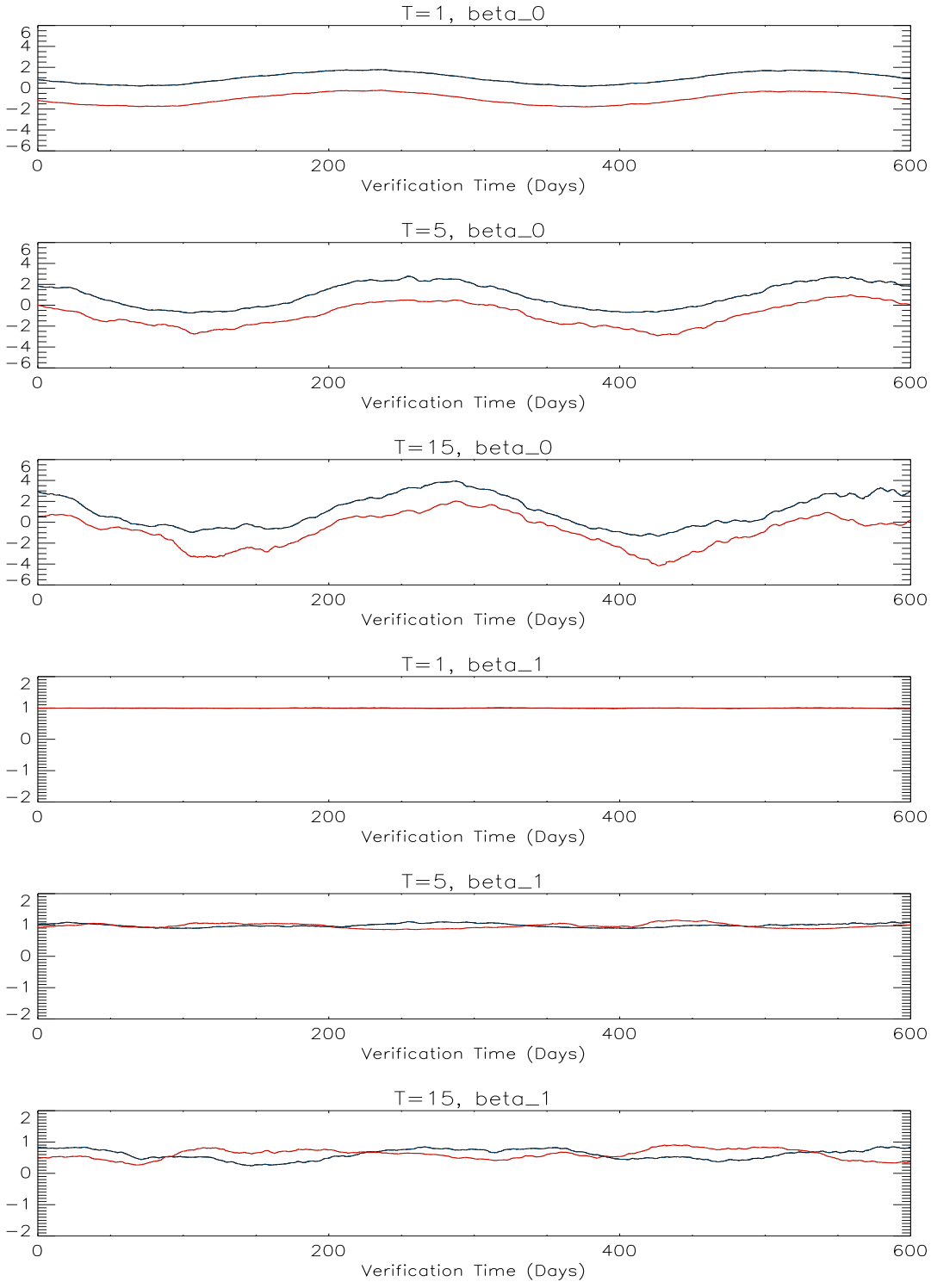


Figure 15: *Bias coefficients computed using reforecast data. The details are as for Fig.11.*

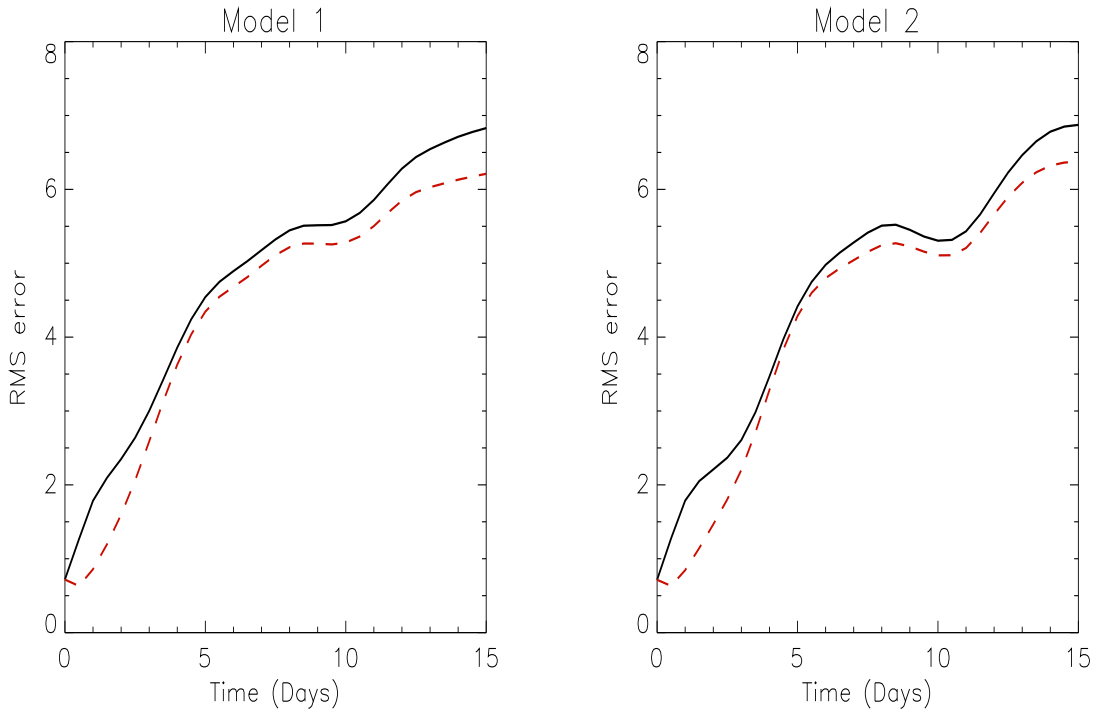


Figure 16: *The RMS errors of the ensemble mean. Black, solid - raw data, red, dashed - bias corrected using reforecasts.*

The RMS errors (Fig. 16) show that the bias correction using reforecast data does improve the ensemble-mean forecast at all lead times. The improvements at long lead times are smaller than those from the moving-window using centred data because the reforecast data only allows the seasonal component of the bias to be reduced, whereas the moving-window also reduces the flow-dependent component of the bias. It is surprising that although it appears, for example from Fig. 11, that the flow-dependent components dominate the biases, the correction of the seasonal-component can still give significant improvements, given that bias correction is a very delicate procedure.

4 Ensemble combination

Calibration is able to improve ensemble forecasts by reducing model biases and correcting the ensemble spread. We now investigate whether ensemble forecasts can be improved further through the use of a combination of ensembles from different models. We investigate both the benefits of a simple combination where all the models are weighted equally and of sophisticated combinations using model-dependent weights. Three different methods to compute the weights are tested: skill-based, multiple-regression based and Bayesian model averaging (BMA).

For a fair comparison, all the multi-model ensembles have the same number of ensemble members as the single-model ensembles. This ensures that the impact due to the increase in the ensemble-size is removed.

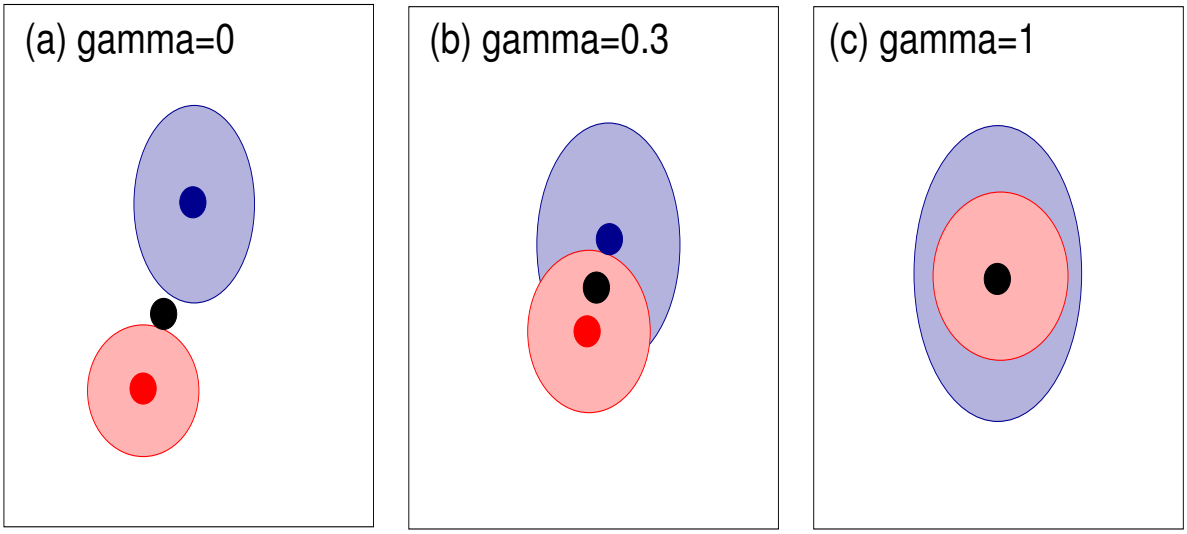


Figure 17: *Diagram of the combination of two ensemble pdfs. The combination methods show (a) the original ensemble members, (b) the single-model ensemble means are nudged towards the multi-model mean, (c) the single-model ensemble means are the same as the multi-model ensemble mean.*

4.1 Simple combination

In the simple combination ensemble, the multi-model ensemble is formed by the union of the bias-corrected ensemble members from the two models. The multi-model ensemble mean is then given by

$$\bar{z}_{MM} = \sum_k w_k \bar{z}_k \tag{20}$$

where the weights are the same for all the models, $w_k = 1/K$. There are then two forms of variance adjustment that are applied. The first is to adjust the between-model variance and the second is to adjust the within-model variance.

To adjust the between-model variance, the new single-model ensemble means are given by

$$\bar{z}_k^{\text{new}} = (1-\gamma)\bar{z}_k + \gamma \bar{z}_{MM} \tag{21}$$

where γ is a scalar parameter that determines how much the calibrated single-model ensemble mean $\bar{z}_k$ should be nudged towards the multi-model ensemble mean $\bar{z}_{MM}$. The corresponding between-model variance is given by

$$var_{between} = \frac{1}{K}\sum_k \left((1-\gamma)(\bar{z}_k - \bar{z}_{MM})\right)^2 \tag{22}$$

and the new mean is associated with the new single-model ensemble members which are given by

$$(z_k^i)^{\text{new}} = z_k^i + \gamma(\bar{z}_{MM} - \bar{z}_k) \text{ for } i = 1, \ldots, N_k \tag{23}$$

The within-model variance of the single-model ensembles is then adjusted so that the multi-model ensemble variance is representative of the mean-square-error of the multi-model ensemble mean. Equation (9), showing how the multi-model ensemble variance is related to the single-model ensemble variances, can be used to derive the appropriate variance adjustment parameters given in equation (8). Using the mean-square-error of the multi-model ensemble mean, MSE_{MM}, the single-model ensemble

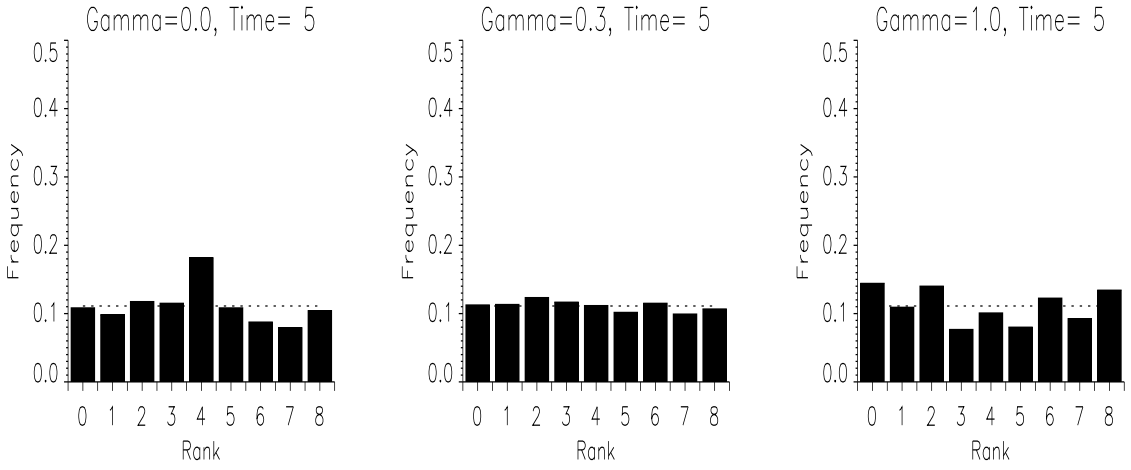


Figure 18: *Rank histograms for the simple multi-model ensemble, with different parameters for γ in equation (21), at time 5 days.*

variance var_k, and between-model variance, $var_{between}$, the variance-adjustment parameter is given by

$$\alpha_k^2 = \frac{1}{var_k w_k M} \left(MSE_{MM} - var_{between} \right) \tag{24}$$

where

$$MSE_M M = \frac{1}{T} \sum_j (\bar{z}_{MM}(t_j) - x^t(t_j))^2 \tag{25}$$

$$var_k = \frac{1}{T} \sum_j \frac{1}{N_k} \sum_i (\bar{z}_k(t_j) - z_k^i(t_j))^2 \tag{26}$$

$$var_b etween = \frac{1}{T} \sum_j \frac{1}{K} \sum_k ((1-\gamma)(\bar{z}_{MM} - \bar{z}_k))^2 \tag{27}$$

This assumes that MSE_{MM} is larger than $var_{between}$. This is not guaranteed, for example if the ensemble-mean is perfect, and hence an extra condition should be added that if $MSE_{MM} < var_{between}$, the α_k^2 are given by a specified positive number close to zero.

Figure 17 illustrates the impact of different values of γ. The red, solid circle shows the ensemble mean from model 1, with the associated ensemble spread shown by the surrounding shaded circle. Similarly, the blue solid circle and surrounding shaded circle show the ensemble mean and spread from model 2. The black circle shows the multi-model ensemble mean. In Fig. 17(a), with $\gamma = 0$, the multi-model ensemble members are equivalent to the single-model ensemble members. The single-model ensemble means are not close to each other because the model-biases have not been removed completely. This results in a bi-modal pdf. In Fig. 17(c), with $\gamma = 1$, the single-model ensemble means are adjusted so that they are the same as the multi-model ensemble mean. This results in a uni-modal pdf, but more variance inflation is required. In Fig. 17(b), with $\gamma = 0.3$, the single-model ensemble means are nudged towards the multi-model ensemble mean. This results in a pdf that is less bi-modal than in (a), but that still contains some of the effects of different models.

Figure 18 shows rank histograms for the Lorenz-model ensembles on day 10 for different values of γ. With $\gamma = 0$, the ensemble is overdispersive, with frequencies concentrated in the middle. This is

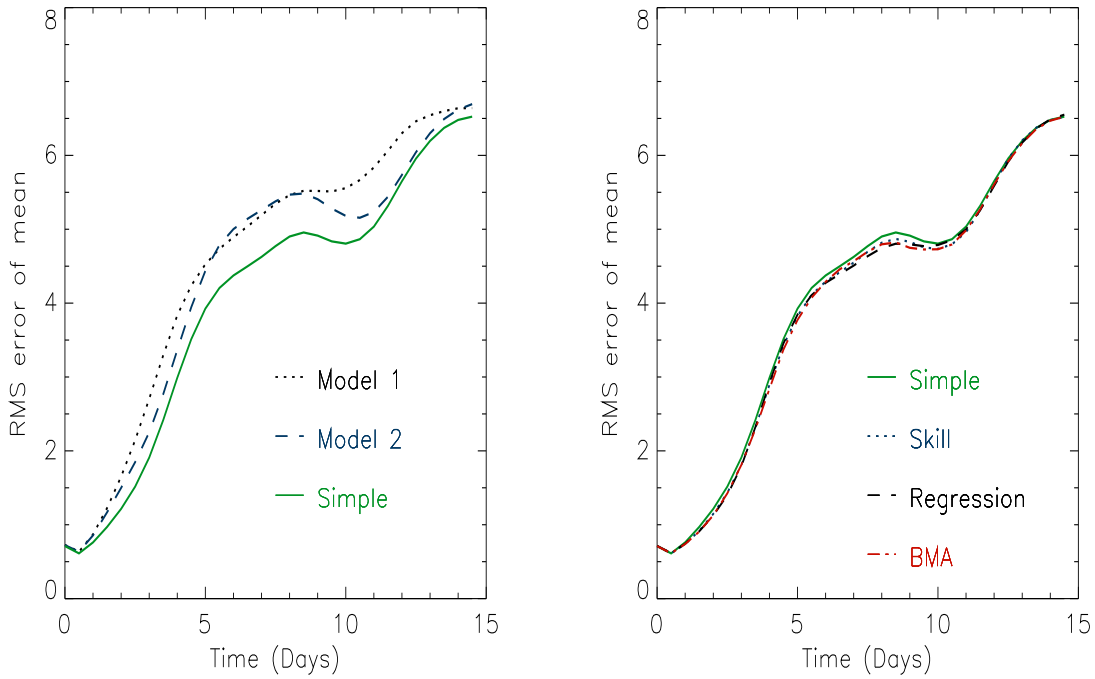


Figure 19: *The RMS errors of the ensemble means for different ensembles. The left panel shows the RMS errors for the two calibrated single models, and for the simple combination multi-model ensemble. The right panel shows the RMS errors for multi-model ensembles with different weight estimation methods.*

because the ensemble pdf is bimodal. With $\gamma = 1$, the ensemble is underdispersive, with frequencies concentrated in the extreme ranks. This is because some of the impact of the different models has been removed so that the ensemble members are very similar to each other. With $\gamma = 0.3$, the rank histogram is closer to the ideal horizontal line and hence this value is used to form the simple-combination.

The left hand panel in Fig.19 shows that the simple combination gives a reduction in the RMS errors of the ensemble-mean, particularly at long lead times. The left hand panels in Fig. 20 shows that a simple combination gives a significant improvement in the resolution with a small reduction in the reliability.

This improvement is due to the averaging of the probabilities, in a similar way that a more accurate prediction is obtained from an ensemble mean than from a single forecast (see Appendix A). When the ensembles are averaged together, the probabilities that give extreme values (close to 0 or 1) but are inaccurate are averaged together, to give probabilities that are closer to 0.5. This means that there are less extreme probabilities in the averaged-ensemble, but the proportion of extreme probabilities that are accurate is much higher than in the single-model ensemble, hence giving a better resolution. There is likely to be little change in the reliability, because at the extremes, the probabilities are closer the observed frequencies, whereas at the mid-range, the probabilities are further from the observed frequencies. This change in the probabilities is shown in the top panels of Fig. 21. The sharpness diagram shows that the simple combination is slightly less sharp than model 2. The values of q near to 0 and 1 are more accurate than for model 2, but the values at the mid-range near to 0.5 are less accurate.

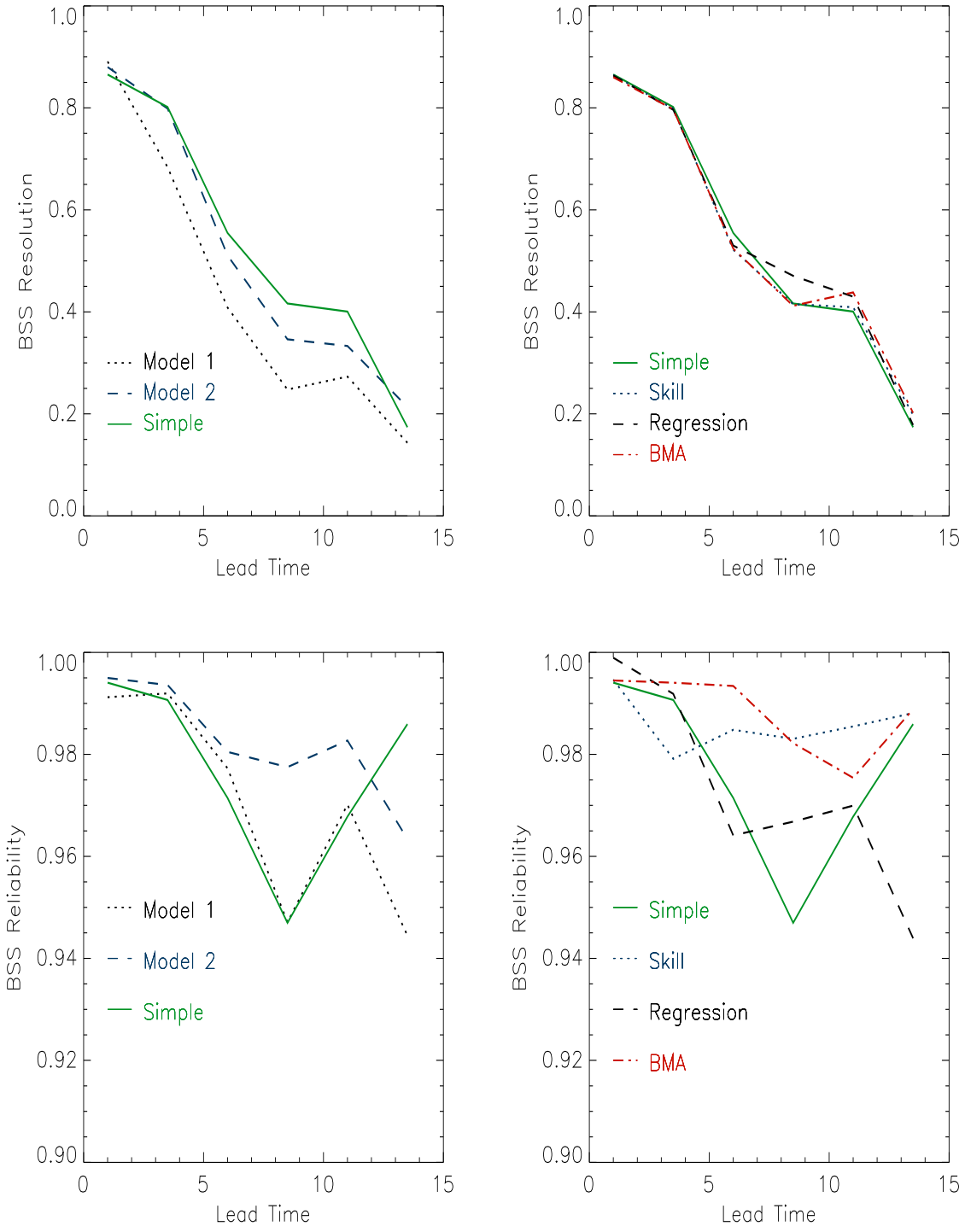


Figure 20: *Brier skill scores for different ensembles. The left panel shows the skill scores for the two calibrated single models, and for the simple combination multi-model ensemble. The right panel shows the skill scores for multi-model ensembles with different weight estimation methods.*

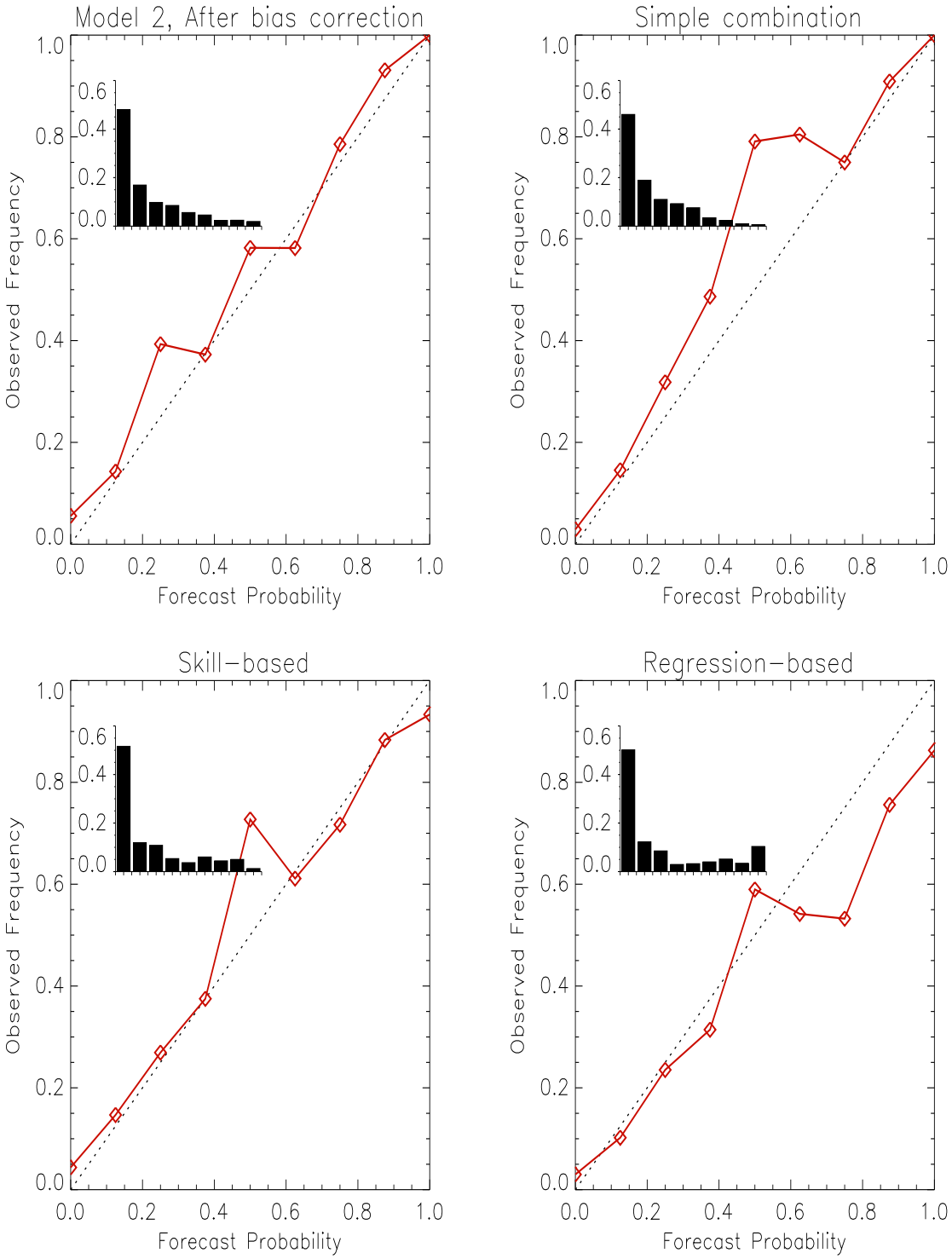


Figure 21: *Reliability diagrams for different ensembles: calibrated single-model, simple multi-model, and multi-models using skill-based and regression-based weight estimation methods. The details are as Fig. 8.*

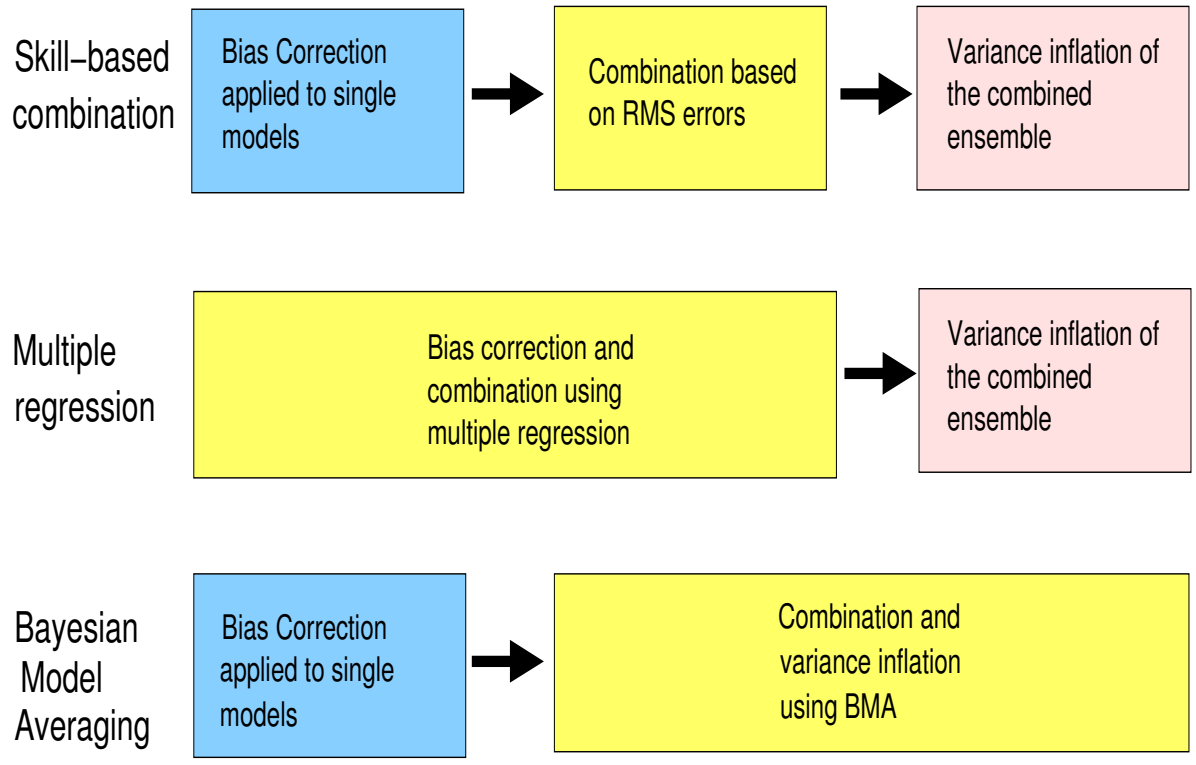


Figure 22: *Illustration of the steps involved in the three combination methods.*

4.2 Weighted combination: skill-based

We now describe three different methods to compute the model-dependent weights. The aim is to give the largest weight to the model that is most likely to be the best model in a given situation. The weights w_k vary between 0 and 1 with 1 being the most skillfull.

The first method is a so-called skill-based method where the weights are dependent on a measure of forecast skill. This method is commonly used in generating so-called consesus forecasts, which are a weighted average of forecasts from different centres (e.g. Woodcock and Engel, 2005). Here, we choose to use the inverse mean-square-error of the ensemble-mean as the measure of skill. The model weights for the M models are computed as:

$$w_k = \frac{(MSE_k)^{-q}}{\sum_{l=1}^{M}(MSE_l)^{-q}} \tag{28}$$

for model numbers $k = 1, \ldots, M$, where MSE_k is the mean-square-error of the bias-corrected ensemble mean for model k over the calibration window and the power q is specified as $q = 1$. In this method, variance-adjustment is applied to the between-model variance using $\gamma = 0.3$.

The steps involved in the skill-based combination method are summarized in Fig. 22. Bias correction is first applied to the single-model ensembles, the model-dependent weights are then calculated from the bias-corrected ensemble-means and variance adjustment is applied to the members from each single-model ensemble.

4.3 Weighted combination: multiple-regression based

The second method to compute the weights is based on multiple-regression. In multiple-regression, we find the linear combination of models that gives the best least-squares fit to the data. In the

multiple-regression based method, instead of applying bias correction to the two single models and then combining, multiple regression is used to combine and bias correct simultaneously, as shown in Fig. 22. This method has previously been applied to many real data sets (e.g. Krishnamurti et al., 1999; Kharin and Zwiers, 2002; Doblas-Reyes et al., 2005; Stephenson et al., 2005).

The method consists of computing the coefficients $\beta_i,\ i = 0, \ldots, M$ such that the new multi-model ensemble mean

$$\bar{z}_{MM} = \beta_0 + \sum_{k=1}^{K} \beta_k \bar{x}_k \tag{29}$$

gives a better least-squares fit to the reference (true) data set y, where $\bar{x}_k$ is the mean of the ensemble from model k. For more than two models, the coefficients can be found by solving a simple linear matrix equation, but for two models the coefficients can be written explicitly. The coefficients β_1 and β_2 act to weight the model to extract the true signal whilst filtering the model error and give more weight to the model that has the highest signal-to-noise ratio. If the coefficient is large, then the model has a relatively small amount of noise and gives a good fit to the observations. In contrast, if the coefficient is close to zero, the model has a relatively large amount of error and it is better to remain close to the climate mean. The coefficient β_0 acts to reduce the overall model bias, ensuring that the time-mean of the combined model is the same as the climate mean of the observations.

In the multiple-regression approach the bias coefficients are used to form a new ensemble mean. The single-model ensemble perturbations are added to the ensemble mean to create the ensemble members. In this way, $\gamma = 1$ in the variance adjustment of the between-model variance. That is, the new single-model ensemble means are identical to the multi-model ensemble mean.

As the bias coefficients β_1 and β_2 do not necessarily add up to one, they can not be used as the model-dependent weights directly. Instead, as suggested by Stefanova and Krishnamurti (2002), the bias coefficients are used to calculate the model-dependent weights

$$w_k = \frac{|\beta_k|^q}{\sum_{l=1}^{M} |\beta_l|^q} \tag{30}$$

where we choose the power as $q = 0.5$. Variance adjustment of the within-model variance is then applied to the members from each single-model ensemble and the model-dependent weights are used to calculate the probabilities from the multi-model ensemble.

4.4 Weighted combination: BMA

The third method to compute the model-dependent weights is the Bayesian-model averaging (BMA) approach proposed by Raftery et al. (2005). In this method, the weights and within-model variance are computed simultaneously, as shown in Fig.22.

4.4.1 BMA applied to individual ensemble members

In the standard BMA method, as proposed by Raftery et al. (2005), the pdf of the forecast variable is formed by the sum of the pdfs associated with each ensemble member, x_i.

$$p(x|\theta) = \sum_i w_i f_i(x|x_i, \sigma_i^2) \tag{31}$$

where f_i is a normal distribution with mean given by the ensemble member x_i and a specified standard deviation σ_i. θ is the set of parameters to be estimated. These parameters are the weights w_i and the variances σ_i^2.

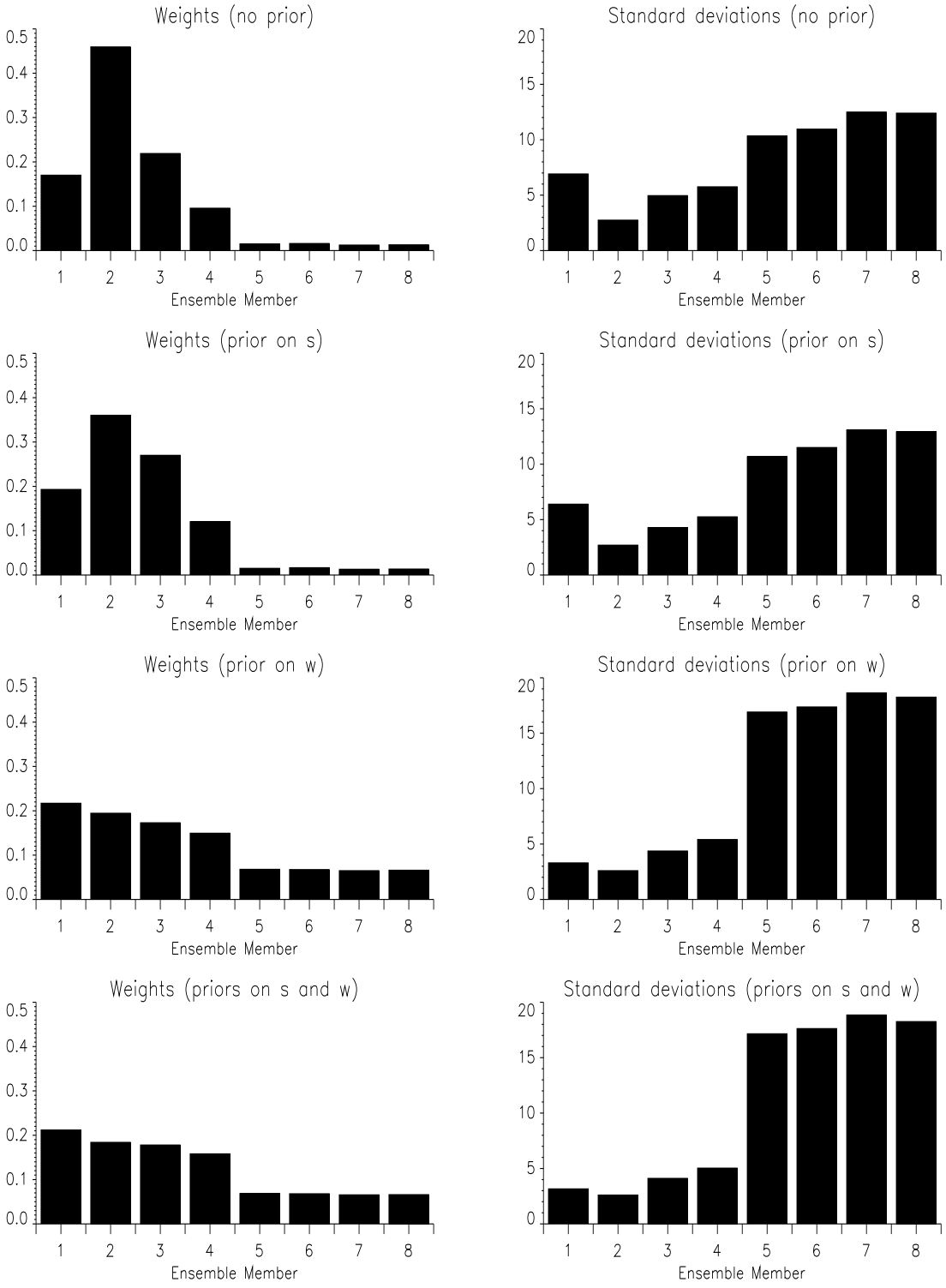


Figure 23: *Weights and standard deviations for each ensemble member produced by BMA applied to each ensemble member. Ensemble members 1-4 are from model 1 whilst members 5-8 are from model 2.*

Raftery et al. (2005) propose that the optimal parameters are found based on maximum likelihood estimation. This formulation is equivalent to the mixture density problem, where the aim is to determine the distributions of populations, given data that contains a mixture of the different populations. In such problems, it is common to find the maximum likelihood estimate using the EM algorithm, an overview of which may be found in Redner and Walker (1984).

The log-likelihood function, to be maximized, is given by:

$$l = log \prod_j \sum_i w_i f_i(y(t_j)|x_i(t_j), \sigma_i^2) \tag{32}$$

$$= \sum_{j=1}^{J} log \sum_i w_i f_i(y(t_j)|x_i(t_j), \sigma_i^2) \tag{33}$$

where $y(t_j)$ is the calibration data at time t_j.

Redner and Walker (1984) note that this problem can be ill-conditioned so that the solution is sensitive to errors in the data, giving inaccurate estimates of the weights and variances and a slow converge rate. This ill-conditioning is a particular problem when the mixtures are not well separated and there is a small sample size. A further problem is that the log-likelihood function is degenerate, as the normal distribution becomes singular when σ is zero. These problems are similar to the ill-conditioning of quadratic minimization in variational data assimilation and can be overcome by adding prior information to the log-likelihood function to give a MAP (Maximum A Posteriori) estimate rather than a MLE (Maximum Likelihood Estimate). Following Fraley and Raftery (2005) and Smith (2004), it is assumed that the variances have an inverse-gamma distribution and that the weights have a beta distribution,

$$p^{\sigma}(\sigma_i^2) = (\sigma_i^2)^{-\frac{e+3}{2}} exp(-\frac{\nu^2}{2\sigma_i^2}) \tag{34}$$

$$p^{w}(w_i) = w_i^a (1 - w_i)^a \tag{35}$$

These priors are zero when $\sigma_i^2 = 0$ and $w_i = 0$ or 1, and positive otherwise. In the following, the parameters ν, e and a are specified as $\nu = 20$, $e = 3$ and $a = 5$. The equations to calculate the optimal estimates, for both the MLE and MAP, are given in Appendix B.

The top panels in Fig. 23 shows the weights and standard deviations produced by the standard BMA method without prior information. Although ensemble members $1 - 4$ are from the same model, and hence have the same levels of skill, they have different weights. Differences in the samples for each ensemble member are emphasised by the ill-conditioning of the problem, leading to different weights.

The lower three panels in Fig. 23 show the weights and standard deviations produced by the same method but with different contraints. The addition of the prior on the variances has no significant impact on either the weights or the variances, probably because the variances without the prior are not close to zero. The addition of the prior on the weights does, however, have a significant impact. The estimates are less sensitive to sampling errors and hence the weights for ensemble members from the same model are now of a similar size. There is a smaller difference between the weights of the two models than without the prior information, but there is also a larger difference in the standard deviations, so that model 2, with smaller weights, is associated with larger standard deviations than without the prior information.

4.4.2 BMA applied to single-model ensemble means

In the application of BMA to atmospheric ensembles, Raftery et al. (2005) assumed that each ensemble member is treated independently and that each ensemble member is the mean of a normal distribution.

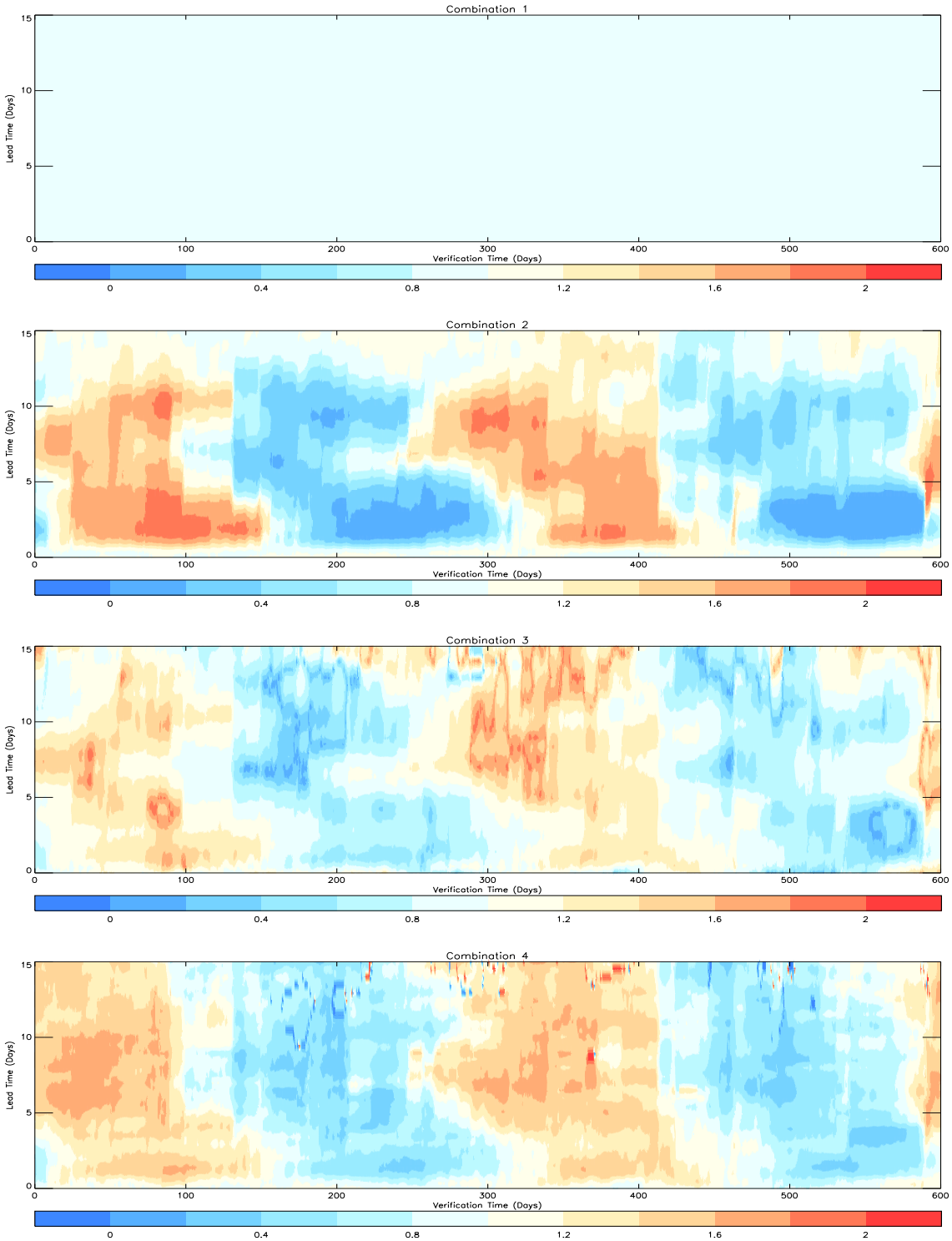


Figure 24: *The weights given to model 1 over a 2 year period (600 days) starting with $\theta = \pi$. The top panel is for the simple combination, the second panel is for the RMS-based combination, the third panel is for the multiple-regression based combination and the bottom panel is for the BMA combination.*

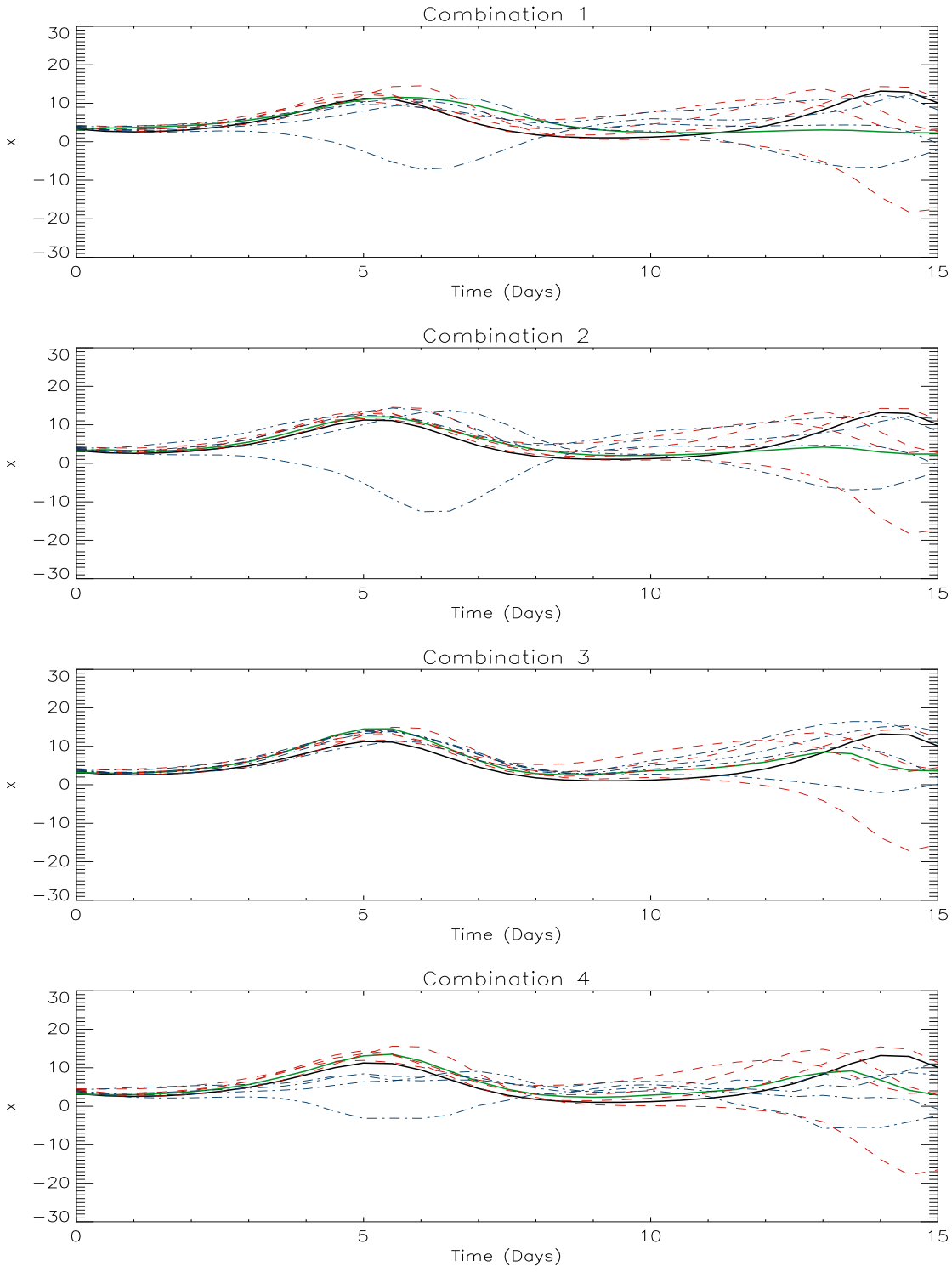


Figure 25: *Ensemble members using calibration data with a realistic moving-window. The top panel is for the simple combination, the second panel is for the RMS-based combination, the third panel is for the multiple-regression based combination and the bottom panel is for the BMA combination. The black solid line shows the true state, the green solid line shows the multi-model deterministic forecast, and the red and blue dashed lines show the ensemble members from models 1 and 2 respectively.*

This is similar to the technique of ensemble dressing, thus leading to smaller sampling errors. Here, we apply the EM-algorithm to calculate the weights and within-model variance adjustment factors, α_k^2, in equation (4) instead of treating each ensemble member independently. This allows us to make a fair comparison with the skill-based and multiple-regression based methods. It also allows for simple verification tools to be applied.

As the weights and variance inflation factors are calculated simultaneously, it is not possible to adjust the between-model variance so that in equation (21), $\gamma = 0$.

4.5 Model-dependent weights results

We now compare the three different methods to estimate the model-dependent weights. An example of ensemble members for the different combination methods are illustrated in Fig.25. Although similar, there are some noticeable differences. It is hard to distinguish the models for the multiple-regression method because this used $\gamma = 1$, whereas there is a clear distinction for the BMA method which used $\gamma = 0$. The deterministic forecasts for all three weight-estimation methods are closer to the truth than for the simple combination because the weights are able to determine which model is more accurate. Finally, the ensemble members for model 2 in the BMA method are less smooth in time. This is likely due to the effect of sampling errors on the estimate of the standard deviations.

The right panel in Fig.19 shows that the model-dependent weights have a slight positive impact on the RMS errors of the ensemble mean. The right panels of Fig.20 shows that there is little change in the resolution, with a slight improvement at long lead times. There is a general improvement in the reliability. By giving the ensemble members different weights, the multi-model ensemble draws closer to the single-model ensemble that is more reliable. In the reliability tables (Fig. 21, the observed frequencies are now increased at the values $q = 0$ and $q = 1$ giving an ensemble that is more sharp. Although there is a slight loss in accuracy at these extremes, there is an improvement at the mid-ranges, giving an overall improvement in both the reliability and resolution.

The weights that are calculated by each method are shown in Fig.24. All three weight-calculation methods give similar estimates for the weights, and clearly show the seasonal variations that were introduced into the models. There are some differences in the weights, and the BMA method has some noisy variations, especially at longer lead times.

5 Conclusions and discussion

Idealized experiments with the Lorenz 1963 model have been used to investigate methods to calibrate and combine multi-model ensembles. Despite the simplicity of the models, the results replicate some of the findings from experiments with operational models and give new findings that will guide multi-model ensemble development.

The first result is that bias correction of the first and second moments, and variance inflation applied to the ensemble spread improves the reliability of the ensemble. Bias correction aims to calibrate the ensemble. It can not add information and hence it can only improve the reliability of the ensemble and not the resolution. Although the bias correction improves the bias of the ensemble it does not necessarily result in improvements in the RMS errors at long lead times. The RMS errors at longer lead times can be improved if the calibration window is centred on the forecast to be calibrated but this is not possible in a realistic situation. The errors at longer lead times can also be improved through the use of reforecast data. However the improvements are not as good as for the centred moving-window because the reforecast data is not able to correct the flow-dependent biases.

The second result is that the simple combination of ensembles gives a significant improvement in the RMS errors of the ensemble mean and the resolution of the ensemble, even though the multi-model ensemble has the same number of ensemble members as the single-models. This result was also found by Hagedorn et al. (2005). This result is not unexpected. In the same way that the mean of an ensemble of deterministic forecasts gives a reduction in the RMS errors, the mean of an ensemble of probability forecasts gives a reduction in the Brier skill scores. Hagedorn et al. (2005) also conclude that 'error cancellation and non-linearity of the diagnostics are the main reasons for the multi-model performance being superior to the average single-model performance'.

The third result is that the model-dependent weights improve the reliability of the ensemble and result in a decrease of the RMS errors. In both this study and in Stefanova and Krishnamurti (2002), the weights are are used in the probability calculations and not to create new ensemble members. This formulation is also consistent with the general framework of combining averaging pdfs. Doblas-Reyes et al. (2005) had concluded that there seemed to be little improvement from the addition of model-dependent weights. However, they had combined the ensemble members by creating weighted averages of the ensemble members from each model rather than defining weights for the calculation of the probabilities.

Three methods to combine the ensembles were tested: a skill-based measure, a multiple-regression based method and Bayesian model averaging. All three methods to estimate weights performed equally well as measured by the Brier skill scores and RMS errors. The weights that were estimated were generally the same for all of the methods, although there were some slight variations. Raftery et al. (2005) also noted that their BMA weights reflected the corresponding sizes of the RMS errors. The conclusion seems to be that the critical part to model combination is how the weights are used, rather than how to calculate the weights. Given that the multiple-regression based weights and BMA weights are expensive to calculate but do not bring any added benefit, it is better to use the most simple method to calculate the weights. The RMS-based method tested here works very well, but the use of other measures such as Brier skill scores could also be investigated.

The improvements from the bias-combination were significantly hindered by a poor calibration dataset, but there was still an improvement in the reliability of the ensemble from the use of the weights when using a realistic calibration window. Thus, we can expect that the use of model-dependent weights to also bring benefits to an operational multi-model ensemble.

This study has not considered the impact of spatial and temporal correlations in forecasts. These correlations bring benefits but can also cause problems. The problem is that the correlations between variables are important and so a calibrated and combined ensemble that is to be physically realistic should also contain these correlations. We expect that if these correlations are present in the calibration data, then they will probably also be present in the calibrated forecast. However, this is not necessarily the case and could cause problems if further postprocessing such as cyclone-tracking is to be implemented. The benefits from the correlations is that it may be possible to use data from a region in space and time rather than a specific location to estimate the calibration and combination parameters. This may lead to more robust estimates of the parameters.

A Verification tools

A.1 Ensemble Averaging

One of the most important properties of an ensemble is that the ensemble mean is, on average, more accurate than the deterministic forecast. As noted by Toth et al. (2003), 'once non-linearity becomes

pronounced, the mean of an ensemble that properly describes the case-dependent forecast uncertainty is able to provide a better estimate of the future state of the system than the control forecast'. This result was first described by Leith (1974), and is summarized by Kalnay (2003). We briefly outline the result here also.

The mean square error of the deterministic forecast x_d is given by:

$$MSE_{deterministic} = E[(x^d - x^t)^2] = E[(x^d - x^c)^2] + E[(x^t - x^c)^2] \tag{36}$$

where x^t is the true state, x^c is the climatological forecast and where we have used the assumption that the anomalies between the forecast and the truth are uncorrelated at long lead times.

Similarly, the mean square error of the ensemble-mean for m ensemble members, $\bar{x} = \sum x_i/m$, is given by

$$MSE_{mean} = E[(\bar{x} - x^t)^2] = \frac{1}{m^2} E[(x_i - x^c)^2] + E[(x^t - x^c)^2 \tag{37}$$

by assuming that the anomalies between the ensemble members with each other and with the truth are uncorrelated. Thus, for $m > 1$, the expectation of MSE_{mean} is less than the expectation of $MSE_{deterministic}$.

This proof relies on the fact that the anomalies between the ensemble forecasts are uncorrelated. This is not true at short lead times, where model linearities have dominated the evolution. However, at longer lead times, nonlinearities become more dominant, causing the forecast anomalies to become uncorrelated.

As the Brier score is a mean square error for probabilities, an almost identical proof can be derived for the Brier score and probabilities, showing that we expect a multi-model ensemble to give improved Brier skill scores. As the probabilities are from different models, they may be uncorrelated at short lead times, hence multi-model ensembles may give improvements at shorter lead times.

A.2 Rank Histogram

Another important property of an ensemble is that the ensemble spread is large enough so that the true state could be a member of the ensemble and small enough so that the ensemble skill is better than that of climate. This is known as the consistency condition. A basic verification tool to examine whether an ensemble satisfies the consistency condition is the **rank-histogram** or **Talagrand Diagram** (Wilks (2006), page 317). In the rank histogram, the ensemble members are binned according to whether they are larger or smaller than the observed value. If all the ensemble members are larger than the observation the rank is zero. If all the ensemble members are smaller than the observation the rank is 9 (number of ensemble members +1). For perfect rank-uniformity, all the bars would be the same height.

In this framework, we also include the weighting of the ensemble members into the calculation of the rank histogram. For a particular time instance, the rank of the ensemble is:

$$rank = \sum_k \left(w_k \frac{1}{N_k} \sum_i o_k^i \right) \tag{38}$$

where o_k^i is the binary verification of whether or not the ensemble member x_k^i is larger than the truth x^t.

q_l	$p(q_l, o_l = 1)$	$p(q_l, o_l = 0)$	$p(q_l)$	$p(o_l = 1 \mid q_l)$
0/8	0.038	0.748	0.079	0.048
1/8	0.007	0.005	0.012	0.555
2/8	0.011	0.007	0.018	0.615
3/8	0.005	0.000	0.007	0.800
4/8	0.003	0.001	0.003	1.000
5/8	0.004	0.000	0.005	0.750
6/8	0.011	0.000	0.011	1.000
7/8	0.013	0.000	0.013	1.000
8/8	0.225	0.000	0.225	1.000

Table 1: *Example of a reliability table.* q_l *is the forecast probability of the event occuring,* o_l *is the binary observation of the event (equal to* 1 *if the event occurs and* 0 *if it does not).* $p(q_l, o_l = 1)$ *is the frequency that the probability* q_l *was predicted given that the event oddcured.*

A.3 Reliability Table

Ensemble forecasts are not only useful for providing an ensemble mean that is more accurate than the deterministic forecast, but also an indication of how likely certain forecast events are. The following verification tools are used to verify the probabilities derived from the ensemble and are all based on the **reliability table**, (see Wilks (2006), pg 282) which uses forecast and observation pairs to find the frequencies that an event is observed with the frequencies that an event is predicted. The event is a dichotomous event - it either occurs or does not occur. For example, the precipitation is greater than or equal to 1mm or the precipitation is less than 1mm. Here we consider the probabilities of the following forecast event: is $x > 5$?

The reliability table gives the possible probabilities that could be found by the ensemble, together with the frequencies that these probabilities occur. These frequencies are separated between frequencies when the event was observed ($o_l = 1$) and frequencies when the event was not observed $(o_l = 0)$.

The reliability table is derived by 'binning' the probabilities: for each forecast, calculate the probability, q, that the event occurs (no.ensembles that predict the event / total number of ensembles) and then add that forecast to the relevant bin for the given probability and the observation of whether the event occurs or not (i.e. in columns 2 and 3 of table 1). The other probabilities, $p(q) = p(q, o_l = 1) + p(q, o_l = 0)$ and $p(o_l = 1|q) = p(q, o_l = 1)/p(q)$, can then be derived.

The probabilities are given by

$$q = \sum_k \left(w_k \frac{1}{N_k} \sum_i o_k^i \right) \tag{39}$$

where o_k^i is the binary forecast of whether or not the event is forecast by ensemble member z_k^i ($o_k^i = 1$ means the event is forecast).

The reliability table can be used to create a number of verification graphics. The first is the **reliability diagram**, which plots the observed frequency, $p(o_l = 1|q)$, against the forecast probability, q. Reliability is a measure of statistical accuracy. For perfect reliability, the observed frequency equals the forecast frequency and hence the curve would go straight through the points (0,0) and (1,1). A climatological forecast has perfect reliability. However, a climatological forecast has no sharpness: the ability to distinguish between different forecast events. The **sharpness diagram** is superimposed on the reliability diagram. This is a bar graph of $p(q)$ against q. For perfect sharpness, there would be only two bars, one at $q = 0$ and one at $q = 1$. Sharpness is not a verification method; a prediction system may have perfect sharpness, giving probabilities of either 1 or 0, but if these are not correct

then the system is useless. A measure of both the sharpness and accuracy is known as resolution. A diagram illustrating how the reliability diagram can be interpreted is shown in Fig. 26.

A.4 Brier Skill Scores

The reliability table and diagram can be summarized using scalar quantities knwon as **Brier Skill Scores** (Wilks (2006), page 284). These are scalar values measuring different aspects of the ensemble skill. The Brier Score (BS) can be thought of as the mean-square-error for an ensemble system, and is given by:

$$BS = \frac{1}{T}\sum_{j=1}^{T}(q(t_j) - o(t_j))^2 \tag{40a}$$

where $q(t_j)$ is the forecast probability at time t_j, $o(t_j)$ is the binary observation at time t_j ($o = 1$ if the event is observed and $o = 0$ if the event is not observed). This can be split into two terms : observed and not observed

$$BS = \sum_l p(q_l, o_l = 1)(q_l - 1)^2 + \sum_l p(q_l, o_l = 0)(q_l - 0)^2 \tag{40b}$$

where q_l is the forecast probability at level l and o_l is the binary observation. This is performing an average over the reliability table, where L is the number of bins in the reliability table and l determines the row in the reliability table. A perfect ensembles system has a Brier score of BS=0, and a useless system has BS=1.

The Brier Score can be decomposed to give scores for the reliability and resolution. The basic derivation is as follows. Using the fact that $p(q_l, o_l = 1) = p(o_l = 1|q_l)p(q_l)$ gives

$$BS = \sum_l p(q_l)\left(p(o_l = 1|q_l)(q_l - 1)^2 + p(o_l = 0|q_l)(q_l - 0)^2\right) \tag{41}$$

$$= \sum_l p(q_l)\left(p(o_l = 1|q_l)(q_l - 1)^2 + (1 - p(o_l = 1|q_l))(q_l)^2\right) \tag{42}$$

$$= \sum_l p(q_l)\left(p(o_l = 1|q_l) - q_l\right)^2 + \sum_l p(q_l)p(o_l = 1|q_l)\left(1 - p(o_l = 1|q_l)\right) \tag{43}$$

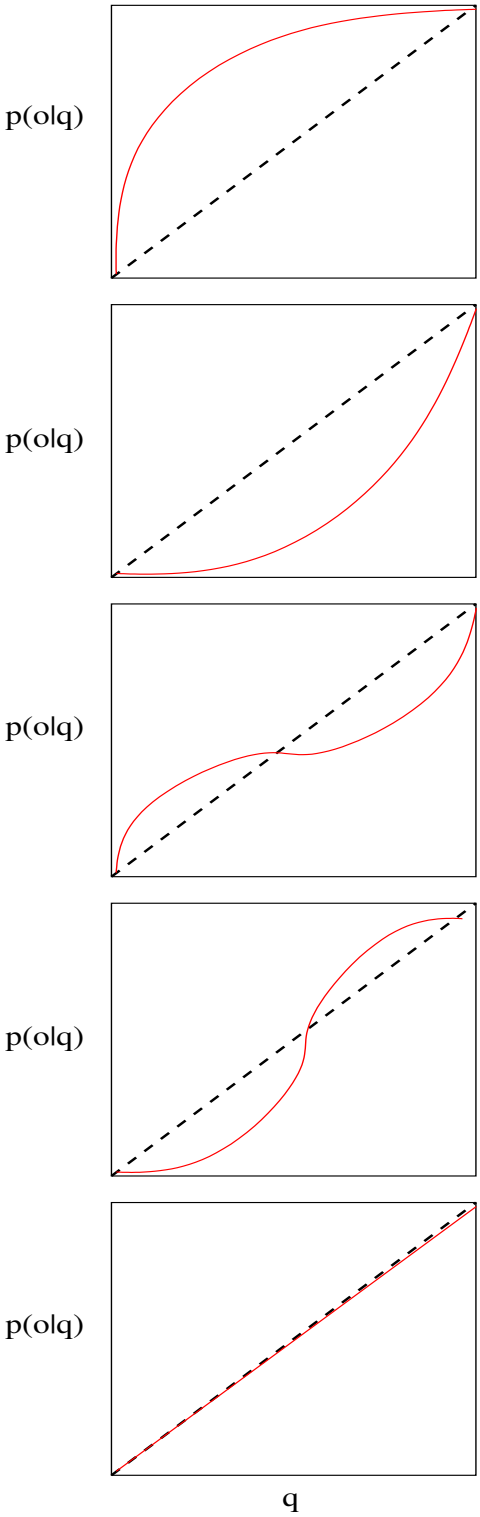
Using the climatological probability of the event occuring, $S = \sum_l p(o_l = 1, q_l)$, gives

$$BS = \sum_l p(q_l)\left(p(o_l = 1|q_l) - q_l\right)^2 - \sum_l p(q_l)(p(o_l = 1|q_l) - S)^2 + S(1 - S) \tag{44}$$

The first two terms are known as the reliability and resolution components,

$$BS_{rel} = \sum_l p(q_l)(q_l - p(o_l = 1|q_l))^2 \tag{45a}$$

$$BS_{res} = \sum_l p(q_l)(p(o_l = 1|q_l) - S)^2 \tag{45b}$$



(a) Underforecasting bias
The model has a negative bias
and so it is very rare to see x>5
Hence the probabilities of this
occuring are very low, when in reality
it occurs much more often.

(b) Overforecasting bias
The model has a positive bias
and so it is very common to see x>5.
Hence the probabilities of this
occuring are very high, when in reality
it occurs much less frequently.

(c) Poor resolution
The system produces probabilities
that are too low when the event is frequent
and probabilities that are too high when the
event is infrequent.

(d) Good resolution
The system produces probabilities
that are too high when the event is frequent
and probabilities that are too low when the
event is infrequent.

(e) Good calibration
The system gives probabilities that
perfectly match the actual frequencies
of events.

Figure 26: *Interpretation of the reliability diagram. Based on Wilks (2006), pg 288.*

Since the BS and BS_{rel} are negatively-oriented measures, (but BS_{res} is positively-oriented) it is more common to use positively-oriented Skill Scores,

$$BSS = 1 - \frac{BS}{BS_{cli}} \tag{46a}$$

$$BSS_{rel} = 1 - \frac{BS_{rel}}{BS_{cli}} \tag{46b}$$

$$BSS_{res} = \frac{BS_{res}}{BS_{cli}} \tag{46c}$$

$$BSS = BSS_{rel} + BSS_{res} - 1 \tag{46d}$$

where BS_{cli} is the Brier Score for climatology. These measures can be interpreted using the reliability diagram (Wilks (2006), page 292). If BSS_{rel} is large, the sum of the squared distances between the points on the reliability curve (solid) and the perfect reliability curve (dashed) is small, for example in Fig. 26, diagrams (c) and (d) have a better reliability than (a) and (b). Similarly, if BSS_{res} is large, the sum of the squared distances between the points on the reliability curve and a horizontal line that intersects the y-axis at the climatological probability, S is large, for example in Fig. 26, diagram (c) has a worse resolution than diagram (d), as the reliability curve is more horizontal.

B The EM algorithm

The main idea of the EM algorithm (Redner and Walker, 1984) is to treat the data y_j as incomplete, and augment the data set with latent variables z, such that $z_{ij} = 1$ if datum y_j belongs to distribution f_i.

In the algorithm we wish to solve:

$$\theta^* = \underset{\theta}{argmax} \log p(x|\theta). \tag{47}$$

This is equivalent to maximizing

$$\log p(x|\theta) = \log p(z, x|\theta) - \log p(z|x, \theta) \tag{48}$$

given the complete augmented data z. By taking the expectation of this equation w.r.t. the conditional distribution of the hidden variables z given the data x and the current estimate of the parameters $\theta^{(n)}$, this gives the equivalent maximization problem

$$\log p(x|\theta) = Q(\theta, \theta^{(n)}) - H(\theta, \theta^{(n)}) \tag{49}$$

where

$$Q(\theta, \theta^{(n)}) = \int \log p(z, x|\theta) p(z|x, \theta^{(n)}) dz \tag{50}$$

$$H(\theta, \theta^{(n)}) = \int \log p(z|\theta, x) p(z|x, \theta^{(n)}) dz \tag{51}$$

Using Jensen's inequality, it can be shown that $H(\theta^{(n+1)}, \theta^{(n)}) \leq H(\theta^{(n)}, \theta^{(n)})$. Hence, $\log p(x|\theta^{(n+1)}) \geq \log p(x|\theta^{(n+1)})$ provided that $Q(\theta^{(n+1)}, \theta^{(n)}) \geq Q(\theta^{(n)}, \theta^{(n)})$. Hence Q provides a lower-bound maximization, and the problem can be solved iteratively as:

$$\theta^{(n+1)} = \underset{\theta}{max} Q(\theta^{(n)}, \theta^{(n)}) \tag{52}$$

B.1 MLE estimate

In the example for BMA without prior information, the Q-function is given by

$$Q = \sum_j \sum_i p_{ij} \log \left(w_i f_i(y_j | x_{ij}, \sigma_i^2) \right) \tag{53}$$

$$= \sum_j \sum_i p_{ij} \left(\log w_i - \log(2\pi\sigma_i) - \frac{1}{2\sigma_i^2}(y_j - x_i)^2 \right) \tag{54}$$

where p_{ij} is the probability that $z_{ij} = 1$. Using calculus of variations, we form the Lagrangian functional $\mathcal{L}$ to incorporate the constraint that the weights w_i must sum to 1.

$$\mathcal{L} = Q + \lambda(\sum_i w_i - 1) \tag{55}$$

The partial derivatives of $\mathcal{L}$, with respect to the variables we wish to optimize, are then given by

$$\frac{\partial \mathcal{L}}{\partial w_i} = \sum_j p_{ij} \frac{1}{w_i} + \lambda \tag{56}$$

$$\frac{\partial \mathcal{L}}{\partial \sigma_i^2} = \sum_j p_{ij} \left(-\frac{1}{\sigma_i} + \frac{(y_j - x_{ij})^2}{\sigma_i^3} \right) \tag{57}$$

Setting these derivatives to zero then gives,

$$w_i = \frac{1}{W} \sum_j p_{ij} \tag{58}$$

$$\sigma_i^2 = \frac{\sum_j p_{ij}(y_j - x_{ij})^2}{\sum_j p_{ij}} \tag{59}$$

where W is specified so that $\sum_i w_i = 1$.

In the E-step, the probabilities that datum y_j was drawn from distribution f_i are calculated

$$p_{ij}^{(n)} = \frac{w_i^{(n)} f_i(y_j | \theta^{(n)})}{\sum_l w_l^{(n)} f_l(y_j | \theta^{(n)})} \tag{60}$$

In the M-step, the new weights and variances that maximize Q are found,

$$w_i^{(n+1)} = \frac{1}{W} \sum_j p_{ij}^{(n)} \tag{61}$$

$$(\sigma_i^2)^{(n+1)} = \frac{\sum_j p_{ij}^{(n)} (y_j - x_{ij})^2}{\sum_j p_{ij}^{(n)}} \tag{62}$$

B.2 MAP estimate

In the example for BMA with prior information given by

$$p^{\sigma}(\sigma_i^2) = (\sigma_i^2)^{-\frac{e+3}{2}} exp(-\frac{\nu^2}{2\sigma_i^2}) \tag{63}$$

$$p^{w}(w_i) = w_i^a (1 - w_i)^a \tag{64}$$

the lagrangian becomes

$$\mathcal{L} = Q + \lambda(\sum_i w_i - 1) + \lambda_\sigma \sum_i logp^\sigma(\sigma_i^2) + \lambda_w \sum_i logp^w(w_i) \tag{65}$$

The partial derivatives are

$$\frac{\partial \mathcal{L}}{\partial w_i} = \sum_j p_{ij} \frac{1}{w_i} + \lambda + \lambda_w a \left(\frac{1}{w_i} - \frac{1}{1 - w_i} \right) \tag{66}$$

$$\frac{\partial \mathcal{L}}{\partial \sigma_i^2} = \sum_j p_{ij} \left(-\frac{1}{\sigma_i} + \frac{(y_j - x_{ij})^2}{\sigma_i^3} \right) + \lambda_\sigma \left(-\frac{e+3}{\sigma_i} + \frac{\nu^2}{\sigma_i^3} \right) \tag{67}$$

which gives new formulae for the M-step,

$$w_i^{(n+1)} = \frac{1}{W} \left(\sum_j p_{ij}^{(n)} + a \left\{ 1 - \frac{w_i^{(n)}}{1 - w_i^{(n)}} \right\} \right) \tag{68}$$

$$(\sigma_i^2)^{(n+1)} = \frac{\nu^2 + \sum_j p_{ij}^{(n)} (y_j - x_{ij})^2}{e + 3 + \sum_j p_{ij}^{(n)}} \tag{69}$$

where W is specified so that $\sum_i w_i = 1$.

References

Doblas-Reyes, F. J., R. Hagedorn, and R. Palmer, 2005: The rational behind the success of multi-model ensembles in seasonal forecasting - ii. calibration and combination. *Tellus*, **57A**, 234–252.

Draper, D., 1995: Assessment and propagation of model uncertainty. *J. R. Statist. Soc. B*, **57**, 45–97.

Evans, R. E., M. H. dn R. J. Graham, and K. R. Mylne, 2000: Joint medium-range ensembles from the Met. Office and ECMWF systems. *Mon. Weather Rev.*, **128**, 3104–3126.

Fraley, C. and A. Raftery, 2005: Bayesian regularization for normal mixture estimation and model-based clustering. Technical Report 486, Department of Statistics, University of Washington.

Hagedorn, R., F. J. Doblas-Reyes, and R. Palmer, 2005: The rational behind the success of multi-model ensembles in seasonal forecasting - i. basic concept. *Tellus*, **57A**, 219–233.

Hamill, T. M., J. S. Whitaker, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Weather Rev.*, **132**, 1434–1447.

Kalnay, E., 2003: *Atmospheric modeling, data assimilation and predictability*. Cambridge University Press, first edition, 341 pp.

Kharin, V. V. and F. W. Zwiers, 2002: Climate predications with multimodel ensembles. *J. Climate*, **15**, 793–799.

Krishnamurti, T., C. Kishtawal, T. LaRow, D. R. Bachiiiochi, C. Williford, S. Gadgil, and S. Surendran, 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, **13**, 4196–4216.

Leith, C., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Weather Rev.*, **102**, 409–418.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20**, 130–141.

Molteni, F., R. Buizza, T. Palmer, and T. Petroliagia, 1996: The ecmwf ensemble prediction system: Methodology and validation. *Q. J. R. Meteorol. Soc.*, **122**, 73–119.

Mylne, K. R. and K. B. Robertson: 2002, Assessment of a multi-centre 'poor persons' ensemble prediction system for short-range use, extended abstract, presented at the AMS Annual Meeting, Orlando.

Palmer, T., A. Alessandri, U. andersen, P. Cantelaube, M. Davey, P. Delecluse, M. Deque, E. Diez, F. J. Doblas-Reyes, H. F. nd R. Graham, S. Gualdi, J.-F. Gueremy, R. Hagedorn, M. Hoshen, N. Keenlyside, M. Latif, A. Lazar, E. Maisonnave, V. Marletto, A. P. Morse, B. Orfila, P. Rogel, J.-M. Terres, and M. C. Thomson, 2004: Development of a European multimodel ensemlbe system for seaonal-to-interannual prediction (DEMETER). *Bull. Amer. Meteor. Soc.*, 853–872.

Palmer, T. N., 1993: Extended-range atmospheric prediction and the Lorenz model. *Bull. Amer. Meteor. Soc.*, **74**, 49–65.

— 2002: The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Q. J. R. Meteorol. Soc.*, **128**, 747–774.

Raftery, A., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Weather Rev.*, **133**, 1155–1174.

Redner, R. and H. Walker, 1984: Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, **26**, 195–239.

Smith, T., 2004: A spatial mixture model of innovation diffusion. *Geographical Analysis*, **36**, 119–145.

Stanski, H. R. and L. J. Wilson, 1989: Survey of common verification methods in meteorology. WMO/TD 358, WMO.

Stefanova, L. and T. N. Krishnamurti, 2002: Interpretation of seasonal climate forecast using Brier skill score, the Florida state university superensembles and the AMIP I dataset. *J. Climate*, **15**, 537–544.

Stephenson, D., C. Coelho, F. J. Doblas-Reyes, and M. Balmaseda, 2005: Forecast assimilation: a unified framework for the combination of multi-model weather and climate predictions. *Tellus*, **57A**, 253–264.

Toth, Z. and E. Kalnay, 1993: Ensemble forecasting at NMC: the generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu: 2003, Probability and ensemble forecasts. *Forecast verification: a practitioner's guide in atmospheric science.*, I. Joliffe and D. Stephenson, eds., John Wiley and Sons, 137–163.

Whitaker, J. S. and T. M. Hamill, 2002: Ensemble data assimilation without perturbed observations. *Mon. Weather Rev.*, **130**, 1913–1924.

Wilks, D. S., 2006: *Statistical methods in the atmospheric sciences*. Elsevier, second edition, 627 pp.

Woodcock, F. and C. Engel, 2005: Operational consensus forecasts. *Weather and Forecasting*, **20**, 101–111.

Ziehmann, C., 2000: Comparison of a single-model eps with a multi-model ensemble consisting of a few operational models. *Tellus*, **52A**, 280–299.